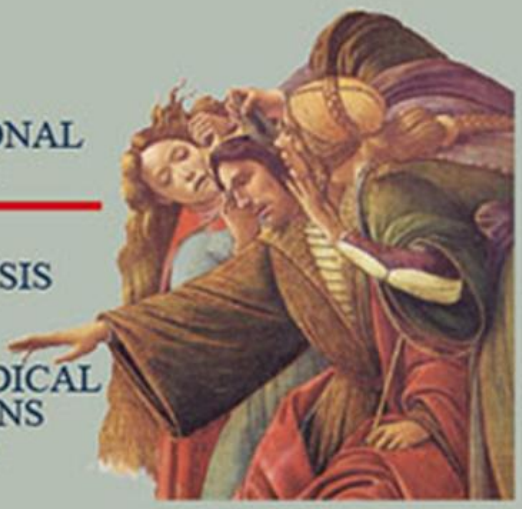
Università degli Studi di Firenze

DET
Dipartimento di Elettronica
e Telecomunicazioni

7th
INTERNATIONAL
WORKSHOP

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

August 25-27, 2011
Firenze, Italy



PROCEEDINGS

FIRENZE UNIVERSITY PRESS

MODELS AND ANALYSIS OF VOCAL EMISSIONS FOR BIOMEDICAL APPLICATIONS

7th INTERNATIONAL WORKSHOP

August 25-27, 2011
Firenze, Italy

Edited by
Claudia Manfredi

Firenze University Press
2011

Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2011.
(Proceedings and report ; 77)

http://digital.casalini.it/9788866550112

ISBN 978-88-6655-009-9 (print)
ISBN 978-88-6655-011-2 (online)

612.78 (ed. 20)
Voce – Patologia medica

Cover: designed by CdC, Firenze, Italy.

Università degli Studi di Firenze
Firenze University Press
Borgo Albizi, 28, 50122 Firenze, Italy
http://www.fupress.com/

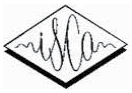
Printed in Italy

The event is sponsored and supported by:

Università degli Studi di Firenze

ENTE CRF - Ente Cassa di Risparmio di Firenze

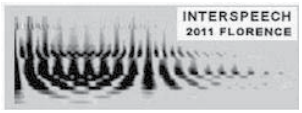
ISCA - International Speech and Communication Association



A.I.I.M.B. - Associazione Italiana di Ingegneria Medica e Biologica

AIVS - Associazione Italiana Scienze della Voce

The MAVEBA 2011 workshop will be held as satellite event of INTERSPEECH 2011



CONTENTS

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Session III: Signal analysis

Session IV: Signal analysis

Session VI: Devices

FOREWORD

It is my pleasure to welcome you to the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA 2011, and to Firenze, Italy, the city that is known especially for its art and architecture, its cultural heritage and its importance in the Middle Ages and that is considered the birthplace of the Renaissance.

Last several decades have witnessed the continued revolution and scientific advances in technology development in the field of biomedical engineering and medicine for better health care. The MAVEBA 2011 presents a broad spectrum of research papers from all over the world, emerging from multidisciplinary areas such as electronic engineering, medicine, mechanical engineering, physics, computational sciences, to focus on the challenges of developing future knowledge in the field of voice analysis, demonstrating the international appeal of this interdisciplinary field that is indeed greater then "models and analysis" as the conference title suggests. However, the MAVEBA roots are strong, demonstrating its commitment to the ideal of bringing together specialists, practitioners and all those interested in aspects of voice care.

I am proud to bring you a great synergy of expertise from internationally renowned scientists and pioneers in medicine and biomedical engineering through keynote speakers and special sessions. The scientific program is targeted to provide tangible benefits and networking opportunities to current researchers; it also provides the best environment for younger scientists and students to learn about future prospects and professional development activities.

The Workshop comprises two sessions centred on internationally renowned keynote speakers:

Prof. Franco Fussi, ASL Audiological Phoniatric Centre, Ravenna, Italy, and Prof. Sergio Daniel Cano Ortiz, University of Oriente, Santiago de Cuba, Cuba, concerning singing voice and newborn infant cry, respectively.

Three Special Sessions encompass highly timely and relevant topics:

- "Computational and experimental vocal fold modelling", organized by Prof. S. Thomson, Provo, Brigham Young University, Utah (USA) and Prof. C. Brücker, TU Bergakademie, Freiberg (D)
- "Innovative ways for acoustic analysis of non-quasi-periodic voices", organized by P. Dejonckere, Cath. Univ. Leuven, Fed. Inst. Occup. Diseases, Brussels (B), and Utrecht Univ. (NL).
- "Acoustic analysis of Parkinsonian speech: issues methods and applications", organized by Prof. S. Sapir, University of Haifa, Haifa (IL)

Moreover, the Workshop hosts six sessions on both traditional and hot topics in the field: Obstructive sleep apnoea, Imaging, Signal analysis, Professional voice, Devices. Again high level scientists will present most recent research and results.

I hope that you will be intellectually stimulated and challenged by the MAVEBA 2011 scientific program while Firenze provides you with an enjoying and charming atmosphere.

In addition to the outstanding scientific programme, I hope that you will find time to explore Firenze and its magnificent natural surroundings.

With my best wishes for a productive meeting

Claudia Manfredi
Conference Chair
Università degli Studi di Firenze

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Infant cry

Invited Speaker and introduction:

Prof. S.D. Cano Ortiz

CRY-BASED NEWBORN DIAGNOSIS OF CNS DISEASES AND SPEECH DEVELOPMENTAL ASPECTS: SOFTWARE AND HARDWARE TOOLS, CRY DATABASES, METHODOLOGIES

S.D.Cano Ortiz[1]

[1]Universidad de Oriente, Faculty of Electrical Engineering, Stgo de Cuba, Cuba

Abstract: **The analysis of infant cry (also known simply as cry analysis) whose use has become more prevalent due to advances in areas such as digital signal processing, pattern recognition and soft computing, has changed the diagnostic ability of physicians to correctly diagnose new-borns using cry analysis. The present paper examines the results and conclusions reached by the Speech Processing Group (GPV) at the Universidad de Oriente in Cuba, in the area of acoustic characterization and the multilateral and multidisciplinary processing of the cry signal, as well as the design and implementation of a cry-based methodology for the diverse diagnosis in newborns with CNS diseases. The endemic development of tools for the acquisition and processing of the cry signal, the coordination of multidisciplinary research team-work (with areas like the logaopedics, phoniatrics, linguistics, neuro-physiology, etc) to carry out a rigorous research schedule, the implementation and testing of hybrid cry classifiers, the development of unprecedented cry-based methodology for diagnosis in newborns affected by CNS diseases (with hypoxia background) as well as the induction of Web-based technology in order to create skills for people involved in the introduction and application of the cry-based methodology in hospital settings, are properly commented.**

Keywords: **cry analysis, signal processing, diagnosis tool**

I. INTRODUCTION

In the last years the research priority of the GPV has been the cry analysis oriented for new-born diagnose, that which responds to a strategic target: only basic and applied investigations within the Cuban Health Care system (CHCS). have the maximum priority, because of which it holds the highest financial support in the country. In Cuba the health care system is full free and it is focused to the primary and preventive attention.

The facilities offered by *our university-hospital-community scheme* within the CHCS empowers the development of multi-disciplinary research project like the one we are leading today: *cry analysis oriented to newborn diagnosis.*

In fact we know much about the cries of both healthy and sick infants, but a reliable investigation procedure, which can be used for clinical purposes, has yet not been developed. During these years we have been managing the hypothesis that it is really possible the development of clinical routines that, supported in the cry analysis, make easier the differential and preventive diagnose of illnesses concerned with the Central Nervous System (CNS) in new-born.

The formulation of this question represents the central scientific problem that has inspired our investigative efforts during last 20 years

II. METHODS

The Scandinavian experience of the 60´s has represented an indispensable piece in our work, masterfully exposed in their research work entitled ***A Spectrographic Studies on Infant Cry*** under the orientation and guide of Prof.Wasz Hocker. [1-3]

Our investigation is also theoretically supported by the Golub´s theory as well as the Theory of Adult Speech production developed by Fant and Flanagan. [4-6]

In order to obtain all the acoustic characteristics and parameters of the cry signal several DSP techniques and algorithms that have been proved to be effective for adult speech signals were properly implemented, among them the most representative were FFT analysis, linear prediction, cepstral analysis, short time analysis and adaptive filtering,.

In the study of the main cry attributes and parameters, we have tried to be enough wider and comprehensive possible, moving around all the diverse representation domains like: time domain (energy, zero-crossings rates, cross- correlation), in the frequency domain (estimation of the fundamental frequency, resonant frequencies or formants, etc.), subjective characterization (soundness, biphonation, vibrato, glottal pulse, tenseness, bifurcation, etc.), all of them estimated by digital spectrograms.
Several approaches have been received special attention in order to compute the fundamental frequency and formants, where the SIFT algorithm has prevailed as well as the cepstral analysis.

As interesting observation we can say that the use of the Mel-scale frequency cepstral coefficients (MFCC´s), something like for adult speech, have fulfilled very satisfactory results in the cry analysis field.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

A. Control Groups

Table 1 presents the 6 groups of control study held during the investigation.

TABLE 1
GROUPS OF CONTROL STUDY

Healthy Control Groups (2)	*Normal childbirth*
	Caesarean
Pathological Control Groups (4)	*Hypoxia*
	Hypoxia with aggravating factors
	Hiperbilirrubinemia
	CIUR (Intra-uterine Growth Delay)

A pain cry was induced by a standardized stimulus: a "heelstick". The infants were positional supine and flat in an open cribs and were not crying at the time of the cry stimulus. Each 12-second cry signal was recorded by a SONY CFS-210 tape recorder with a flat frequency response from 40 to 20,000 Hz. A hand held PHILIPS SBC-3040 microphone was used to allow recording of the cry at a distance of approximately 17 cms of the baby's mouth. The recording was done by the researchers at the Southern Maternity Hospital of Santiago de Cuba. Then the recorded cries were digitized by a high-speed microcomputer (with PCVOX acquisition system). With those original cry data and their corresponding clinical profiles the BDLLanto database v1.0 was initially constructed.

The soft and hard supporting was partially developed by the researchers of GPV (65%)

Thanks to the collaboration of foreign institutions as the UPM[1] of Spain and the VUB[2] of Belgium the acquisition of equipments needed for our investigations has been really possible, letting us avoid the crude effects of US embargo in this kind of research project.

Our own tools and procedures have also been validated in a standard Kay Elemetrics station located in the Dept. of Neurolinguístics of the VUB Hospital in Brussels, thanks to the appreciable collaboration of Prof. Jan Raes.

B. Cry Analysis and Classification

In this process several processing alternatives have been applied to the acoustic cry analysis such as: *auditory analysis, time-frequency analysis of the cry signal, spectrographic analysis, digital signal processing (DSP) techniques*, all of them empowered by the development of computers and new information technologies. To the classical approach of threshold-based infant cry analysis (that means to extract relevant diagnostic information from the threshold behavior of acoustic cry parameters [1-3] [5-7], we recently added soft-computing approaches like combinatorial logic, connectionist model, the genetic-neural and hybrid systems.

The necessity to process automatically high volumes of information presupposes the implementation of a well structured process with its defined integral blocks as it is shown in Fig. 1.

Starting from the signal data acquisition it follows a block that processes the cry signal in order to obtain the feature vectors or attributes. Next a block that includes reduction of dimensionality defining the final vectors to be present at the input of classifier, finally it appears the classification block that defines the ownership to one class or another

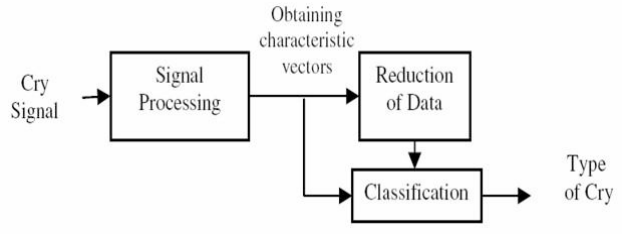


Fig. 1 Automatic Infant Cry Recognition (AICR) process

This is a key part of the process since here the acoustic features which will be the classifier input are properly estimated. In our experience, to the primary attributes concerned with the characterization of the vocal tract and excitation source (F0, Fi)[3] the well known MFCC´s are added with very good yields. [8-9]

The classification stage assumes 2 phases: *training* and *tests*. In this stage it has been working intensely in the last three years leading to the new focuses as follow:

- Logical-combinatorial analysis
- Connectionist model
- Neuro-evolutive analysis
- Neuro-fuzzy approach

Logical-combinatorial analysis: this is related to the logical-combinatorial approach of Pattern Recognition whose essential idea is to establish the analogy in which an object may resemble another, but it might not be in its entirety. This approach is an alternative to the statistical approach, regularly applied in medical research.

Connectionist model: these systems are known as connectionist models or Artificial Neural Networks (ANN), due to the resemblance its processing has with

[1] UPM: Universidad Politécnica de Madrid
[2] VUB: Vrije Universiteit Brüssel

[3] F0-fundamental frequency, F_i-formant frequencies

the form of processing of the human nervous system. This approach has been used in the classification of infant crying under several scenarios.

Neuro-evolutive analyis: this approach is a recent detour used to select the best features of the crying input vectors, which are used to train a classification system based on neural networks. To make this selection, Evolutionary Strategies techniques are applied.

Neuro-fuzzy approach: a work titled "Type-2 Fuzzy Sets Applied to Pattern Matching for the Classification of Cries of Infants under Neurological Risk" was presented in [10] consisting in a pattern recognition algorithm for the classification of infant cries. But very recently we tried to classify infant cry by compressing the original signal, instead of reducing the vectors once they were analyzed. Here the reduction method uses Fuzzy Relational Product (FRP) to compress the information inside a feature vector, building with this a compressed matrix that will help us recognize two kinds of pathologies in infants: *Hypoxia* and *Deafness*. This algorithm uses codebooks to build a small relational matrix that represents an original vector [10-11]. Thus the Fuzzy approach becomes on a viable alternative for the cry classification oriented to the newborn diagnose.

III. Results and Discussion

A. A Cry-based methodology for New-born diagnosis

Several experiments were done in order to test new procedures for cry signal classification. The performances from neural-evolutive approach and FRP deal with the highest classification rates reported by technical literature [10]. Moreover the percents of efficiency with FRP are lightly higher than those obtained recently by other approaches within the soft computing field. It is also very interesting to note the improvement in the fuzzy-classification performance when this approach is properly combined with the FRP-compression techniques At the same time a ANN-Threshold classifier also reached high classification rate in [9]. That emergency of hybrid cry classifiers is opening a new way for automatic cry classification and opportunity for cry-based methodology with diagnostic purposes.

The proposal of *a cry-based methodology* incorporates all our experience in the infant cry analysis including the new focuses of cry classification. In Fig. 2 there is a brief description of it.

A first block of signal acquisition follows the established recording protocol and the filling of the model sheet 01 (with all the clinical profile of the new-born)..

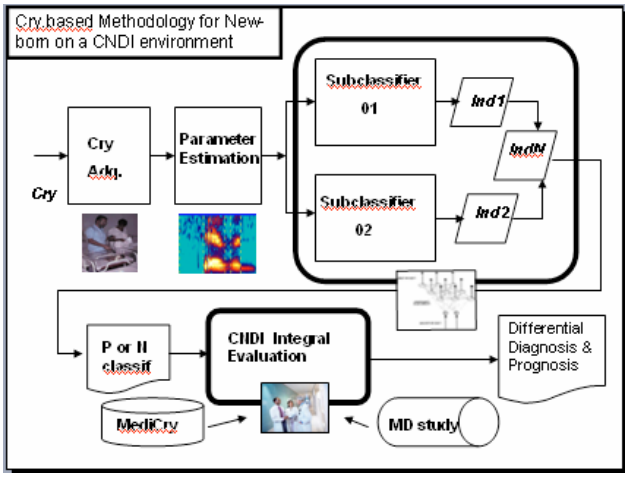


Fig. 2 A proposal of Cry-based methodology for new-born diagnosis

A second block assumes the estimation of all the cry attributes (time domain, frequency domain and qualitative features)

Next the classification block establishes 2 pre-classifiers (one threshold-based classification with F0 as feature and one hybrid pre-classifier with a combination of threshold-based classification and classification with RNA).

Each pre-classifier grants a classification sub index, those classification sub indexes serve as an input to the final block where the normality index is computed, which finally locates the cry input signal in one of the 2 classes (P-pathologic or N-normal).

This information upgrades the MediCry corpus and at the same time is available for the multidisciplinary staff within the Consultation of the Infant Neuro-developmental Outcomes (CNDI), that it is in charge to evaluate integrally the clinical information (given by the physical recognition of baby, reading from specialized equipment and clinical tests), the result of the multilateral cry processing (not only the result from the cry classifier but also from the study of each acoustic parameter in the digital spectrogram). Through the CryTrainer v1.0 a better preparation of the CNDI staff members is fulfilled, just to assume a multidisciplinary focus in the use of all the cry information concerned with the methodology.

Actually the cry-based methodology is being introduced at the Southern Maternity Hospital of Santiago de Cuba for its pre-evaluation protocol.

B. Other Tools for Supporting the methodology

Recently collateral results has been developed by the GPV consistent on soft tools supported by Web technology, which should facilitate the visibility of our work, the constant exchange of experiences among specialists of cry community as well as practical

application of the cry-based methodology and its integral validation. They are:

- *MediCry*: a MySQL database which holds all the clinical information and cry corpus concerned with the research project
- *CryTrainer*: a web-based trainer dedicated to train cry researchers in order to read digital spectrogram of cry signal
- *WebSA on Cry*: A Web-based platform for Infant Cry Analysis for a collaborative environment that lets get an effective exchange of research experiences and information sharing among the researchers within cry community

B. Work in progress

We are directing our efforts to the Classification of Cries of Infants under Neurological Risk and High Risk Born. For this task we will try to get a large infant cry corpus with as many types of samples and from as many pathologies as possible. Among the samples we will try to build crying histories of medical interesting cases.

We are also doing research on how to automatically identify qualitative acoustic features in the crying signals, which until now are only found by visual observation. Among them are features like vibratos, glides, glottal rolls, melody, etc. Next we will try to associate them with some physical status or pathology.

V. CONCLUSION

The GPV during last 20 years has been hardly working in the field of cry analysis looking for clues and potentials features to be used in early detection of CNS diseases and differential diagnosis in new-born. Its main results has been synthetically commented showing that now we are closest than ever to reach cry classifiers able to support new-born diagnosis in clinical routine.

In the future the final product we visualize for the automatic cry-based methodology is conceived as a system with several embedded stages to be called a ***N****euro-physiological* ***E****valuation* ***S****ystem for* ***N****ew-born* (SENF). It will include a block of cry data acquisition, a classification block, on-line access to digital clinical information supported in Web technology, a report generator, Web connection with MediCry database as well as a dynamic friendly-user interface available for the CNDI staff.

REFERENCES

[1] S. Karelitz, V.R. Fisichelli, (1962) "The cry thresholds of normal infants and those with brain damage". J. Pediat. 61, 679-685.

[2] O. Wasz-Höckert, J. Lind, V.Vuorenkoski, T. Partanen, E. Valanne (1968) "The infant cry a spectrographic and auditory analysis". Clinics in Devel. Medicine. 29.

[3] K. Michelsson (1982) “Sound Spectrographic cry analysyis of normal and abnormal newborns”. Folia Phoniatria, 28, pp. 161-173.

[4] H.L.Golub, M.J. Corwin (1985) "A physioacoustic model of the infant cry". In Lester, B. M., Boukydis, C. F. Z., (eds). Infant Crying: Theoretical and Research Perspectives. N. York, Plenum Press. 59-82.

[5] H.L.Golub, M.J. Corwin (1982) "Infant cry: a clue to diagnosis". Pediatrics. 69 (2), 197-201.

[6] B.M.Lester (1984) "A biosocial model of infant crying". In Lipsitt, L. P., (ed). Advances in Infancy Research. N. York, Academic Press. 167-212.

[7] A. Fort, C. Manfredi, C. "Acoustic analysis of newborn infant cry signals". Med. Eng. Phys. 20, (6), 432-442, Sep, 1998.

[8] O.F. Reyes, S.D. Cano-Ortiz, C.A. Reyes (2008) "Validation of the Cry Unit As Primary Element for Cry Analysis Using An Evolutionary-Neural Approach", 9no Encuentro Internacional Mexicano de Ciencias de la Computacion ENC 2008, oct 2008, Universidad Autonoma de Baja California, Mexicali, Mexico.

[9] S.D. Cano-Ortiz, D.I. Escobedo, I. Suaste, T. Ekkel, C.A. Reyes Garcia: (2006) "A Combined Classifier of Cry Units with New Acoustic Attributes. " In Fco.Martinez Trinidad et al Progress in Pattern Recognition, Image Análisis and Applications, CIARP 2006, Lecture Notes in Computer Science LNCS 4225, pp 416-425, Springer Verlag Berlin Heidelberg 2006, Cancun, Nov 13-17 2006, México.

[10] K. Santiago-Sánchez, C.A. Reyes García, P. Gómez-Gil, “Type-2 Fuzzy Sets Applied to Pattern Matching for the Classification of Cries of Infants under Neurological Risk”, in Lecture Notes in Computer Sciences (LNCS) 5754, edited by De-Shuang Huang, et al, Springer, Berlin, 2009, ICIC 2009, pp. 201-210, ISBN: 978-3-642-04069-6, ISSN: 0302-9743.

[11] K Hirota, W. Pedrycz. Fuzzy Relational Compression. IEEE Transactions os Systems, man, and cybernetics,Part B: Cybernetics, Vol. 29, No. 3, June, pp. 1-9. (1999)

STUDY OF CRY PATTERNS IN INFANTS AT HIGH RISK FOR AUTISM

S. Orlandi[1], L. Bocchi[1], C. Manfredi[1], M. Puopolo[2], A. Guzzetta[3] S. Vicari[4] and M.L. Scattoni[2]
[1] Dept. of Electronics & Telecommunications, Università degli Studi di Firenze, Firenze, Italy
[2]Department of Cell Biology and Neurosciences, Istituto Superiore di Sanità, Rome, Italy
[3] Department of Developmental Neuroscience, Stella Maris Scientific Institute, Pisa, Italy
[4]Department of Neuroscience, Children's Hospital Bambino Gesù, Rome, Italy

Abstract: **Autism Spectrum Disorders (ASD) show an increasing prevalence in children, and are often undiagnosed up to the third year of life. Analysis of cry patterns appears a promising approach for allowing an early ASD detection and diagnosis. In this work we compare the main acoustic parameters collected by recording infant crying in control subjects with the parameters obtained in high risk subjects, namely infant siblings of children with ASDs. Results confirm previous finding obtained using home video recordings, and indicate a weaker coupling between fundamental frequency and first resonance frequencies in high risk subjects.**
Keywords: **Autism Spectrum Disorders, cry analysis, fundamental frequency, resonance frequencies.**

I. INTRODUCTION

Cry constitutes the first communication channel available to newborns for fulfilling their needs and attracting attention of the caregivers. Hearing a crying baby produces, in the adults, a reaction aimed at activating parental caretaking, ensuring newborn survival and comfort.

Cry involves activation of both the newborn and listener neural system, increasing the reciprocal attention level. It is produced when the newborn perceives a negative stimulus, from an internal or external source, and it involves a coordinated effort of several brain regions, mainly brainstem and limbic system.

For this reason, cry can be candidate as an early sign of potential problems and pathologies involving the neural system, and it should be included and analyzed during the evaluation of newborn state.

Recently, there has been a large interest in the analysis of cry features in children with Autism Spectrum Disorders (ASD), mainly because of the major role played by brainstem and limbic system, both areas compromised in children with ASD [1,2], in the production of infant crying.

Autism is a neurodevelopmental disorder characterized by impairments in social and communication development, and by restricted and repetitive behavior. Typically children are not diagnosed before two years of life despite 50% of parents of children with ASDs report that they suspected a problem before their child was 1 year of age [3], thus a more precocious diagnosis seems possible.

Recent epidemiological studies reported prevalence rates in the general population of 58-67/10.000, suggesting that ASDs affect many families and represent a serious public health problem. Over the years, interventions have focused on enhancing developmental skills and on ways of ameliorating behavioral difficulties by teaching more effective communication skills. There are studies demonstrating that early intensive behavioral intervention initiated at preschool age and sustained for 2-3 years results in substantial improvements for a large subset of ASD children. Gains are found in IQ, language, and educational placement. Although a few pharmacological treatments can reduce some associated symptoms, early behavioral interventions remain the most effective treatment for the symptoms of autism. Thus, early identification of ASDs allows the possibility of early intervention, for the ultimate purpose of optimizing quality of life and functional independence.

Previous studies on the properties of cry in autistic children involved the spectrographic analysis of the sound signal and of the modulation of the acoustic wave, reporting the presence of significant differences between controls and subjects later diagnosed with ASD. More in detail, crying episodes in ASD subjects have shorter duration, less modulation, and lack of regular peaks than crying recorded in control subjects [4]. Moreover, fundamental frequency is lower than in control subjects, and structural properties appear atypical [5,6].

II. METHODS

Acquisition protocol

The project will recruit a set of about 200 control subjects and a set of high risk subjects to be followed prospectively (tentatively, 20 subjects). By allowing accurate and detailed assessments of behavioral measures at fixed time points, prospective studies offer theoretical advantages to detect early modifications of ASD, while avoiding biases.

Presently, no diagnostic tool is available for the early detection of autistic children. Recent advances in early detection research have resulted from prospective studies carried out on high risk infants. We planned to recruit later-born infant siblings of children diagnosed with

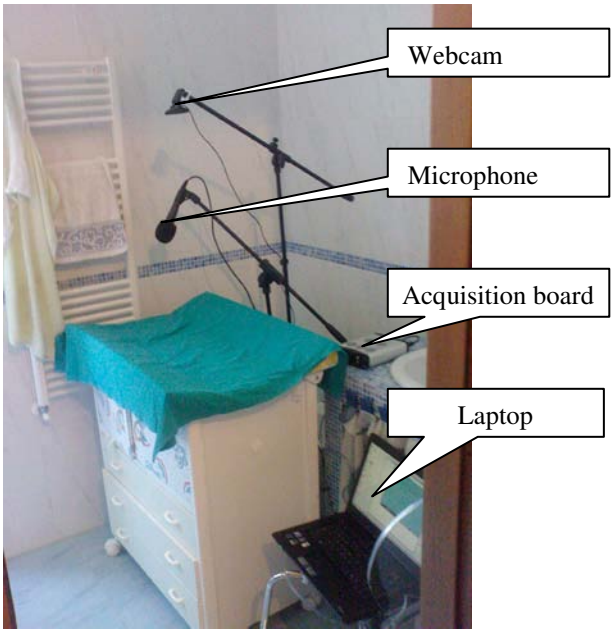


Fig. 1 Acquisition system in a typical setup

ASD. These infant siblings are themselves at especially high risk for an autism or ASD diagnosis [7] and this population is arguably the most clearly defined high risk group available [8,9].

The planned acquisition procedure is totally not invasive, minimizing the ethical issues involved in the recruitment of control subjects and high risk infants.

Each subject is involved in a set of measures, scheduled every six weeks, starting a few days after birth up to the 24th week of life (hence, each one will undergo to five testing sessions). In this work, we present results obtained during the 2nd testing session, recorded approximately during the 8th week of life.

Informed consent has been obtained from all parents. The study protocol has been approved by the local ethical committee (Istituto Superiore di Sanità, IRCCS Fondazione Stella Maris, and IRCCS Pediatric Hospital Bambino Gesù).

The acquisition system has been designed for being used in the patient home, minimizing the discomfort for the involved subjects and the impact of the external environment on children habits. Hence, the basic requirement is the ease in transporting and assembling the system. According to this requirement, the proposed system, shown in Fig. 1, includes a laptop which is connected to an high speed USB video camera (Logitech HD pro webcam C910), able to provide a 1280x1024 pixel video stream and an external audio acquisition device (Tascam US-144-MK2, as the quality of the audio card embedded in the laptop is not adequate to the recording specifications) and a professional microphone (Shure SM58).

Signal processing

In the present work, we focus on the analysis of the audio track, performed using BioVoice, custom software developed in Matlab language. The detail of the algorithms used in BioVoice for processing cry recordings have been already described in [10], however we summarize only part of the elaboration relevant for this study in the following paragraphs. The first processing step aims at detecting each crying episode in order to label it for further elaboration. This allows speeding up the elaboration by removing unnecessary data. Detection of crying episodes is performed using a Voiced/Unvoiced detection procedure. In the next step, ach detected crying episode undergoes a detailed analysis, where the following features are extracted from the signal:

- length and average amplitude of the episode
- fundamental frequency F_0
- resonance frequencies, mainly first and second resonance frequency (F_1 and F_2, respectively)

F_0 estimation and voiced/unvoiced detection

The fundamental frequency, F_0, was estimated with a two-step procedure. Simple inverse filter tracking (SIFT) was applied first [11,12], to signal time windows of short and fixed length. The window length was chosen as $M = 3F_s/F_{min}$, where F_s is the signal sampling frequency and F_{min} is the minimum allowed F_0 value for the signal under consideration (for newborn cry: F_{min} = 150 Hz).

In the second step, F_0 was adaptively estimated inside $[F_l, F_h]$. This allowed for a more precise F_0 estimation. A variable window length for analysis was applied, inversely proportional to the changing F_0. Very short time windows, ranging from 5 to 15 ms, were thus obtained, locally dependent on the signal variability. Over each time window, the signal was band-pass filtered (for newborn cry the range was settled to 150–900 Hz) with the Mexican hat continuous wavelet transform, and the signal periodicity was extracted by means of the average magnitude difference function (AMDF) approach. In case of fast and abrupt F_0 changes, this procedure was shown to increase the robustness of the F_0 estimation, giving enhanced results with respect to standard methods [12].

In order to disregard voiceless parts of the signal, a *voiced/unvoiced decision* (V/UV) was applied. It was based on the approach proposed previously in [13] and was suitably modified for our purposes here. Basically, a signal frame is selected as voiced if voicing evidence, γ_{max}, defined as the amplitude of the autocorrelation function on that frame, is larger than a threshold value. A number of controls made on adjacent frames have been added to ensure continuity of the detected pitch in order to exclude possible wrong V/UV choices [13]. For a newborn cry, it was commonly found that $\gamma_{max} \geq 0.06$.

Resonance frequencies

Even if vowel frequencies cannot be found in newborn cries, resonance frequencies (RFs) reflect important acoustical characteristics of the infant vocal tract. For RFs estimation and tracking, a robust parametric technique is used, obtained by peak picking in the power spectral density (PSD) plot. This was evaluated on the same adaptive time windows as previously described. For PSD estimation, autoregressive (AR) models were used. The model order q varied according to the signal characteristics. The "modified covariance method" was applied, as it was shown to give the best results for the reduction of spectral line splitting and bias of the frequency estimations [12].

The relation $q \cong 0.5F_s$ (in kHz) was found to be optimal for obtaining an enough detailed spectrum, while preventing spectral smoothing and consequent loss of spectral peaks. Co-ordinates of PSD maxima on each time window, as well as their mean and std value on the whole signal, were also evaluated. Thus, details were given about RFs evolution in time as related to energy. The first three RFs, are extracted by the BioVoice software, however in the present work we focused only on the first two of them, F_1 and F_2.

III. RESULTS

Cry episodes detection

The performance of the detector has been assessed by qualitative inspection of the audio signal superimposed on the output of the voice detector. Visual analysis of the resulting waveform indicates a substantially correct extraction of cry episodes.

An example of the detected voice segments is shown in Fig. 2, where a signal frame containing four crying episodes is shown.

Fundamental frequency analysis

A comparison of the characteristics of cry episodes of a control subject and a high risk subject of the same age show some interesting differences. As already described in the introduction, results confirmed the lower frequency range of the fundamental frequency in the high risk newborn with respect to control subjects. Moreover, we studied possible the relationship between F_0 and F_1 during cry episodes. In control cases, we often observed a strong coupling between the two variables. As shown in Fig. 3, a large number of frames shows a linear relation coupling F_0 and F_1: for instance, in the case shown in figure, when F_0 is in the range from 400 to 550Hz, we have found $F_1 = 3\ F_0$ in almost all frames. As shown in the figure, most frames indicate a strong coupling between the two frequencies, accordingly to a linear relation with an angular coefficient which can be expressed as ratio of small numbers (the other alignments which may be seen in the figure correspond to angular coefficients equal to 2+1/4, 2+1/2 and 2+3/4).

By contrast, the scatter plot relating F_0 and F_1 in a high risk subject is usually similar to the one reported in Fig. 4, where the coupling between the two frequencies F_0 and F_1 is weaker.

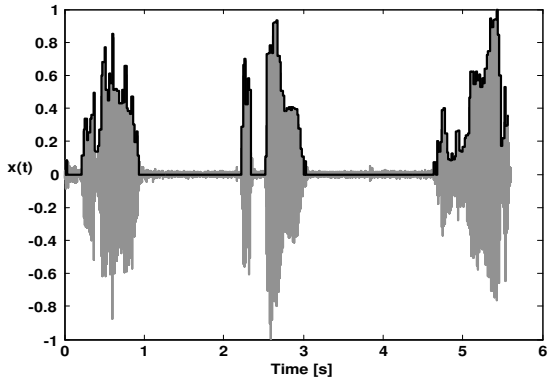


Fig. 2. Detection of cry episodes (black line) and audio signal (gray)

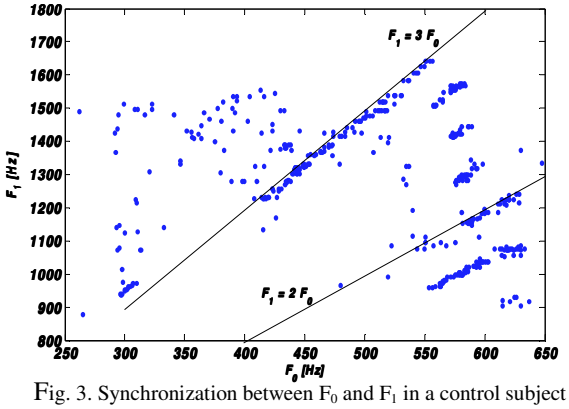


Fig. 3. Synchronization between F_0 and F_1 in a control subject

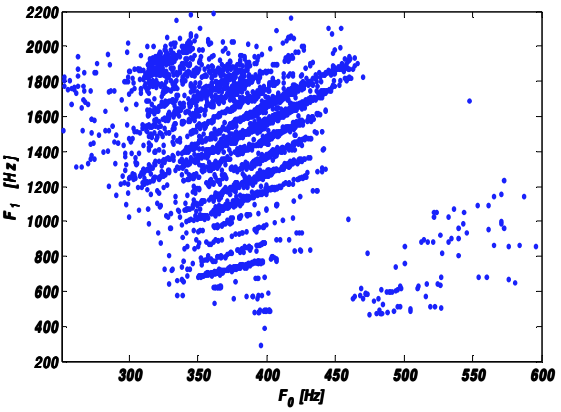


Fig. 4. Synchronization between F_0 and F_1 in a high risk subject

Melody

In each cry episode, the fundamental frequency presents a well-defined trend. Four typical patterns have

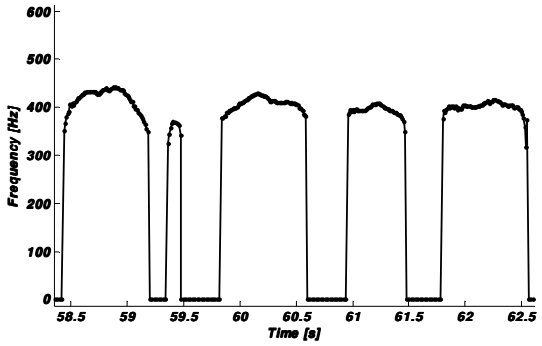


Fig. 5. Typical fundamental frequency trends. First peak is clearly symmetrical, the last one is a plateau. The two central ones are not clearly defined.

been observed in newborns [14,15], namely the symmetrical pattern (frequency rising and falling around a central peak), the rising pattern (frequency peak appears near the end of the episode), the falling pattern (frequency peak appears at the beginning), and the plateau (with an almost constant frequency). A cry recording of a high risk newborn, 2 months old, with samples of symmetric and plateau patterns, is reported in Fig. 5.

IV. Conclusion

Autism Spectrum Disorders (ASD) are often undiagnosed up to the third year of life. Analysis of cry patterns appears a promising approach for allowing its early diagnosis and treatment. In this work main acoustic parameters obtained in control subjects with the parameters found in a small group of high risk subjects are compared. The results of these first experiments indicate that appreciable differences can be found. At present, the sample size, especially as concerns high risk subjects, is too small to assess the statistical significance of these differences. Work is in progress in order to increase the sample size and to define best acoustic parameters suited for an early non-invasive diagnosis of autism spectrum disorders.

Acknoledgement

Supported by the Italian Ministry of Health Grant (GR3), Young Researcher 2008, “Non-invasive tools for early detection of Autism Spectrum Disorders”.

References

[1] Amaral DG, Schumann CM, Nordahl CW. „Neuroanatomy of autism.”. Trends Neurosci. 31(3) pp. 137-45. (2008)

[2] Schulkin, J. “Autism and the amygdala: An endocrine hypothesis”. Brain and Cognition, 65, 87–99 (2008).

[3] E. Werner, G. Dawson, J. Osterling, N. Dinno, “Brief report: recognition of autism spectrum disorder before one year of age: retrospective study based on home videotapes”. Journal of autism developmental disorders, 30, pp. 157-162, 2002.

[4] P. Venuti, G. Esposito, Z. Giusti, “A qualitative analysis of crying and vocal distress in children with autism”, Journal of Intellectual Disability Research, 48, 2004.

[5] P. Venuti , Z. Giusti , F. La Femina , G. Esposito, F. Domini “Qualitative analysis of children’s autistic cry observed by homevideo”, Infant Cry Workshop 8th International Workshop, Padova, 2002.

[6] G. Esposito, P. Venuti, "Typical and atypical expression of distress: A study on cry", Int. J. of Public Health, 1(2), pp. 141-150, 2009.

[7] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, P. Szatmari, “Behavioral manifestations of autism in the first year of life”, Int J Dev Neurosci, 2005 Apr-May, 23(2-3) pp. 143-52.

[8] L. Zwaigenbaum, S. Bryson, C. Lord, S. Rogers, A. Carter, L. Carver, K. Chawarska, J. Constantino, G. Dawson, K. Dobkins, D. Fein, J. Iverson, A. Klin, R. Landa, D. Messinger, S. Ozonoff, M. Sigman, W. Stone, H. Tager-Flusberg, N. Yirmiya , “Clinical assessment and management of toddlers with suspected autism spectrum disorder: insights from studies of high-risk infants”, Pediatrics, 2009 May, 123(5) pp. 1383-91.

[9] Zwaigenbaum L. “Advances in the early detection of autism”, Curr Opin Neurol. 2010 Apr, 23(2) pp. 97-102.

[10] C. Manfredi, L. Bocchi, S. Orlandi, L. Spaccaterra, G.P. Donzelli, “High-resolution cry analysis in preterm newborn infants”, Medical Engineering & Physics, 31(5) pp. 528-532, 2009.

[11] J.D. Markel, “The SIFT algorithm for fundamental frequency estimation”, IEEE Trans Audio Electroac 20, pp. 367–377, 1972.

[12] S.L. Marple, “Digital spectral analysis with applications”, Prentice Hall, Englewood Cliffs, NJ, USA, 1987.

[13] L. M. Van Immerseel, J. P. Martens, “Pitch and voiced/unvoiced determination with an auditory model” J. Acoust. Soc. Am. 91 (6), pp. 3511-3526, 1992.

[14] K. Wermke, W. Mende, “Musical elements in human infants’ cries: in the beginning is the melody”, Musicae Scientiae, 13, pp. 151-173, 2009.

[15] K. Wermke, M. Birr, C. Voelter, W. Shehata-Dieler, A. Jurkutat, P. Wermke, A. Stellzig-Eisenhauer, “Cry Melody in 2-Month-Old Infants With and Without Clefts”, Cleft Palate-Craniofacial Journal, May 2011, 48(3), pp. 321-330.

Session I:
Obstructive sleep apnoea

DETECTION OF OBSTRUCTIVE SLEEP APNEA USING SPEECH SIGNAL ANALYSIS

O. Elisha[1], A. Tarasiuk [2], Y. Zigel [1]
[1]Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2] Sleep-Wake Disorders Unit, Soroka University Medical Center and Department of Physiology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel
orenelis@bgu.ac.il, tarasiuk@bgu.ac.il, yaniv@bgu.ac.il

***Abstract*: Obstructive sleep apnea (OSA) is a prevalent sleep related breathing disorder associated with several anatomical abnormalities of the upper airway. Acoustic parameters of human speech are influenced by properties of the vocal tract, which includes the upper airway. We hypothesize that it is possible to differentiate OSA patients from non-OSA (healthy) subjects by analyzing potential patients' speech signals. Using speaker recognition and signal processing techniques, we designed a system for classifying a given speech signal into one of the two groups. The database for this research was constructed from 92 subjects who were recorded reading a one-minute speech protocol immediately prior to a full polysomnography study; one hundred and three acoustic features were extracted from each signal; seven independent Gaussian mixture models (GMM)-based classifiers were implemented; a fusion process was designed to combine the scores of these classifiers and a validation procedure took place in order to examine the system's performance. Specificity and sensitivity of 91.66% and 91.66% were achieved for the male population; and 88.89% and 85.71% were achieved for female population, respectively. Such a system can be used as a tool for initial screening of potential OSA patients.**

***Keywords*: obstructive sleep apnea, speech signal processing, speaker recognition.**

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a sleep disorder that is caused by obstruction of the upper airway. OSA severity is defined by the number of obstructive apnea and hypopnea events per hour of sleep (apnea hypopnea index – AHI). OSA affects approximately 5% of adults in the western population; a 2- to 3-fold greater risk for men compared to women has been reported [1]. OSA can lead to numerous complications such as hypertension, cardiovascular disorders, and excessive daytime sleepiness [2]. Currently, diagnosis of OSA is conducted in a sleep laboratory where a full polysomnography (PSG) study is performed. PSG is expensive, time consuming, and uncomfortable for the patient.

In earlier studies, researchers found that OSA is associated with several anatomical abnormalities of the upper airway that are unique to this disorder [3]. Acoustic parameters of human speech are affected by the physiological properties of the vocal tract (which includes the upper airway) such as vocal tract structure and soft tissue characteristics. Therefore, it was suggested [4] that acoustic speech parameters of an OSA patient may differ from those of a non-OSA subject (speaker). Our hypothesis is that speech signal properties of OSA patients will be different than those of control (non-OSA) subjects, and that we are able to distinguish between the two groups using a computer-based system that will analyze the subject's voice. The influence of OSA on speech is not yet fully understood but some researchers have tried to classify OSA subjects using speech signals [5] [6]; in both studies one classifier was trained on all speech segments using various acoustic features.

In this study we designed a system that fuses several Gaussian mixture model (GMM)-based classifiers, one for each of the voiced phonemes, using different acoustic features and model parameters. Our primary goal is to use this set of classifiers for initial screening of potential OSA patients that will assist in reducing the number of patients referred to sleep clinics for diagnosis. Our secondary goal is to improve our understanding of the effect of the disorder on speech including investigating the hyper-nasalization degree of the speech signals.

II. METHODS

A. Experiment setup

The test population of this research was constructed from 60 male subjects and 32 female subjects; subjects' age, AHI, and body mass index (BMI) are presented in Table 1. All subjects are patients who were referred to a sleep clinic by different doctors as "potential" OSA patients. All subjects underwent full PSG examination, were diagnosed, and given an AHI by the clinic's medical staff. Each subject was recorded using a digital audio recorder (Handy recorder "H4" by ZOOM) reading a one-minute text protocol in Hebrew, designed by the researchers to emphasize certain elements of speech. In

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

order to avoid over-fitting, the speech data was then divided into two separate databases: design and verification (validation).

Table 1 – The subjects' information

Diagnosis	*Number of subjects*	*AHI average ± Std*	*Age average ± Std*	*BMI average ± Std*
Male				
Healthy	12	4.83 ± 1.79	45.55 ± 13.6	27.66 ± 4.07
OSA	48	28.26 ± 20.17	56.58 ± 13.18	31.2 ± 5.7
Female				
Healthy	14	3.44 ± 2.28	47.23 ± 13.87	28.35 ± 6.73
OSA	18	24.44 ± 17.33	58.65 ± 10.59	33.44 ± 6.07

B. Pre-processing and feature extraction

Each recorded speech signal underwent a pre-processing procedure of down-sampling (to 16 kHz), DC removal, pre-emphasizing, and normalization; followed by manual segmentation of the signal in order to isolate specific phonemes. Using the signals from the vowels (/a/, /e/, /i/, /o/, and /u/) and nasal phonemes (/n/, /m/) alone, the signals were further framed into 30 msec frames. One hundred and three different acoustic features were extracted from each frame. The extracted features can be divided into four groups: time domain features, such as energy, pitch, jitter, and shimmer; spectral features, such as linear predictive coding coefficients (LPC) and their first and second derivatives, formant location and bandwidth, auto regressive moving average (ARMA) coefficients, and other potentially relevant spectral features; cepstrum domain features such as mel-frequency cepstral coefficients (MFCC) and their derivatives; and features for detection of hyper-nasal speech, which will be further elaborated later.

In addition to these "short term features" that were extracted from each frame, another set of features was computed as statistics of some of the short-term features through the entire speech signal, such as average of harmonic to noise ratio and average distance between formants. These "long-term features" represent the stationary position of the vocal tract uttering different vowels [5].

C. Abnormal nasalization degree detection

In [7], the researchers suggested that OSA patients demonstrate an abnormal nasalization degree in their speech. This abnormality is usually caused by a defective velopharyngeal mechanism [8] that may be associated with OSA. Hyper-nasal speech is characterized by amplitude reduction of the first formant, presence of zeros in the spectrum due to coupling of nasal and oral cavities, presence of reinforced harmonics resulting from the sound resonance in the nasal cavity, and a shift of formants [9]. In order to differentiate OSA patients from non-OSA subjects we added three features to the feature extraction process for estimation of hyper-nasalization degree of each given frame.

The first feature is based on a nonlinear operator called Teager energy operator (TEO) [8].

$$\psi\{s[n]\} = s^2[n] - s[n-1]s[n+1] \quad (1)$$

where $s[n]$ is the speech signal in time domain. The TEO can be shown to be sensitive to multi-component signals (such as hyper-nasal speech signals). The extraction of this feature was implemented as follows: each signal was filtered once with a BPF around the first formant and once with LPF, which was set to remove the frequencies that are higher than those of the first formant. TEO was extracted from both signals, and cross correlation between the two outputs was calculated. The assumption is that if there is only one component in the signal (no nasal harmonic near the first formant) the signals will be similar, but in the case of hyper-nasalized speech, the signals will be different.

The second feature proposed is based on using high and low order LPC [9]; where in case of hyper-nasal speech, there will be a large difference between the spectra obtained from these two sets of coefficients. The distance between the LPC sets was calculated by calculating the real LP cepstrum $c(k)$ and finding the geometric distance between the two sets using (2).

$$d = \sum_{k=0}^{\infty}[c_H(k) - c_L(k)]^2 \quad (2)$$

where $c_H(k)$ and $c_L(k)$ are high and low order LP cepstral sequences, respectively.

The third feature is set to detect the spectral flattening associated with hyper-nasalization of a given speech signal. Power spectral density (PSD) was estimated for each frame using Welch's method and standard deviation (STD) was calculated on the PSD between 300 Hz to 2000 Hz [10].

These features were added to previously described features that discriminate between normal and abnormal nasalization degree of speech, such as first formant location and bandwidth, distance between first and second formants, and ARMA coefficient.

D. Feature selection and model estimation

Seven GMM-based classifiers were implemented; one for each of the five vowels, one for the nasal phonemes, and one for "long-term features". Each phoneme-based classifier was trained separately on a different subset of features selected via a sequential forward floating

selection algorithm (SFFS). The most discriminative features for each model were chosen to maximize the performance of the classifier. After designing all seven phoneme-based classifiers and calculating the parameters for an OSA model and a healthy model for each classifier, each subject (of the design data) was tested over all models and scored using log-likelihood ratio and Z normalization [11], getting 7 normalized scores $\Lambda_i(\mathbf{x})$ $(i = 1, \dots, 7)$ – one for each classifier:

$$\Lambda_i(\mathbf{x}) = \frac{\frac{1}{N}\sum_{j=1}^{N} log\left(p(\mathbf{x}_j|\omega_{Hi})\right) - \frac{1}{N}\sum_{j=1}^{N} log\left(p(\mathbf{x}_j|\omega_{Oi})\right) - \mu_O}{\sigma_O} \quad i = 1, \dots, 7 \tag{3}$$

where $p(\mathbf{x}_j|\omega_{Hi})$ and $p(\mathbf{x}_j|\omega_{Oi})$ are the likelihood probabilities of the *j*th feature vector $\mathbf{x}_j$, given the model for healthy subjects and for OSA patients, respectively. μ_O and σ_O are the OSA population's mean and variance, respectively, and *N* is the number of frames.

The significance of each classifier was evaluated by conducting a leave one out (LOO) validation procedure on the design data. A fusion process was performed in order to combine all scores; the fusion process was found on issuing different weight, w_i $(i = 1, \dots, 7)$, to each score based on the significance of the classifiers' results. Classifiers that resulted in total significance of 60% or less were taken out of the final score and the remaining scores were weighted in proportion to their significances; the total of all weights is set to be 1. During the training phase, a threshold was calculated for all classifiers.

E. Validation

A validation procedure took place using the validation data; each subject was tested in a leave one out process, scores were given to the subject for each model, and summed using the previously calculated weight function:

$$\Lambda^w(\mathbf{x}) = \sum_{i=1}^{7} w_i \Lambda_i(\mathbf{x}) \tag{4}$$

The weighted score and the previously calculated threshold were used to decide whether to label each subject as OSA or non-OSA (healthy).

III. Results

Using the design database, the feature selection procedure resulted in a different set of selected features for each classifier; moreover, a different order of GMM was proven more efficient for each different phoneme.

In a recent study conducted in our lab [5], an identical database was used to achieve the same purpose of differentiating OSA from non-OSA (healthy) patients, using a **single** GMM classifier (baseline system, C) for all speech frames; an 8th order GMM model was implemented on a 5-dimension feature space for males and 4th order GMM model was implemented on a different 5-dimension feature space for females. The features in baseline system (C) were selected using the same SFFS procedure and out of the same 100 features described previously in section II, but without the "hyper-nasal" features. In order to evaluate the efficiency of our method of training 7 phoneme-based classifiers separately, and the effect each phoneme has on the final score, we examined our system using the same 5 features selected in [5] for each of the seven classifiers (system B). The results of all 3 systems are presented in Table 2.

Adding the three hyper-nasal speech detection features to the model further improved our result. System A was retrained using all 103 features; results are presented in Table 3.

Table 3 – Results for system A with hyper nasal detection features

Male		
	classified as O	classified as H
true label O	**91.66%**	8.33%
true label H	8.33%	**91.66%**
Female		
	classified as O	classified as H
true label O	**88.89%**	11.22%
true label H	14.29%	**85.71%**

The results of each phoneme-based classifier were fused with the weight function calculated with the design data; this function is presented in Table 4.

Table 2 – Results of 3 different systems (O-OSA, H-healthy)

	System A: 7 phoneme-based classifiers, separate feature selection procedure		**System B:** 7 classifiers, same features		**System C** (baseline system): 1 classifier for all phonemes	
	Male					
	classified as O	classified as H	classified as O	classified as H	classified as O	classified as H
true label O	**85.42%**	14.58%	83%	17%	83%	17%
true label H	16.66%	**83.33%**	33.33%	66.66%	21%	79%
Female						
	classified as O	classified as H	classified as O	classified as H	classified as O	classified as H
true label O	**83.33%**	16.66%	77.77%	22.23%	86%	14%
true label H	14.29%	**85.71%**	21.43%	78.57%	16%	84%

Table 4 – Weight function for each gender

	/a/	/e/	/i/	/o/	/u/	/m/+/n/	Long term
Male	0.16	0.05	0.00	0.00	0.16	0.11	0.52
Female	0.02	0.00	0.00	0.28	0.00	0.70	0.00

IV. Discussion and Conclusion

From Table 2 one can see that the proposed system (A), which offers an optimal feature set for each phoneme and a fusion between phoneme-based classifiers, is superior to the other compared systems (B and C).

For comparison, the results presented in [5] (system C) are 83% specificity and 79% sensitivity (for males). Implementing the same optimal 5 features of system C on the phoneme-based system (B) caused performance degradation to 83% specificity and 66% sensitivity, implying that those five features are not the optimal features for each phoneme.

Adding the hyper-nasal speech detection features to the model further improved the results, increasing specificity and sensitivity to 91.66% and 91.66% for male subjects, and 88.89% and 85.71% for female subjects. These improvements imply a difference in the nasalization properties between OSA and non-OSA groups. In order to further examine this potential discriminating property we trained our system (system A) using only 7 features: 3 hyper nasal features and first and second formants' location and bandwidth. Classification results of 70.8% specificity and 75% sensitivity were achieved, reinforcing the assumption of hyper nasalization in OSA patients' speech.

The procedure of training different classifiers with different feature sets for each phoneme (system A) indeed improved the results; moreover, the weight function and the results of each model led us to conclude that some phonemes (such as /a/ and nasal phonemes) carry more distinguishing information than other phonemes between OSA subjects and healthy subjects.

From the results of this research, it appears that initial screening of potential OSA patients using speech signals is indeed possible.

V. References

[1] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 9, pp. 1217-1239, 2002.

[2] N. J. Douglas, *Harrison's Principles of Internal Medicine*, 17th ed., New York: McGraw-Hill Medical, 2008.

[3] T. M. Davidson, "The great leap forward: the anatomic basis for the acquisition of speech and obstructive sleep apnea," *Sleep Medicine*, vol. 4, no. 3, pp. 185-194, 2003.

[4] T. M. Davidson and J. Sedgh, "The anatomic basis for the acquisition of speech and obstructive sleep apnea: evidence from cephalometric analysis supports the great leap forward hypothesis," *Sleep Medicine*, vol. 6, no. 6, pp. 497-505, 2005.

[5] E. Goldshtein, A. Tarasiuk, and Y. Zigel, "Automatic detection of obstructive sleep apnea using speech signals," *IEEE Trans. on Biomedical Eng*., Vol. 58, No. 5, pp. 1373-82, 2011.

[6] R. F. Pozo, J. L. B. Murillo, L. H. Gómez, E. L. Gonzalo, J. A. Ramírez, and D. T. Toledano, "Assessment of severe apnea through voice analysis, automatic speech, and speaker recognition techniques," *EURASIP Journal on Advances in Signal Processing*, doi:10.1155/2009/982531, 2009.

[7] A.W. Fox, P.K. Monoson and C.D. Morgan, "Speech dysfunction of obstructive sleep apnea. a discriminant analysis of its descriptors", *Chest*, vol. 96 no. 3 pp. 589-595, September 1989.

[8] D.A. Cairns, J.H.L. Hansen, J.F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear Teager energy operator," *Spoken Language,*. ICSLP 96. Proceedings., Fourth International Conference, vol.2, pp.780-783, Oct 1996.

[9] D.K. Rah, Y.I. Ko, and C. Lee, "A noninvasive estimation of hypernasality using a linear predictive model", *Annals of Biomedical Engineering*, Springer Netherlands, vol. 29, no. 7, pp. 587-594, 2001.

[10] T. Pruthi, and C. Y. Espy-wilson, "Acoustic parameters for the automatic detection of vowel nasalization," *in INTERSPEECH2007*, Antwer, Belgium, August 2007.

[11] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

AUTOMATIC DETECTION OF SNORING EVENTS USING GAUSSIAN MIXTURE MODELS

E. Dafna,[1] A. Tarasiuk,[2] Y. Zigel[1]
[1]Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer–Sheva, Israel
[2]Sleep-Wake Disorders Unit, Soroka University Medical Center, and Department of Physiology, Faculty of Health Sciences, Ben-Gurion University of the Negev, Israel

Abstract: **In this work, an automatic snore detection system of acoustic snoring signals has been designed. Its purpose is to assist an alternative non-invasive method for diagnosing obstructive sleep apnea (OSA) based on acoustic signal processing. The detector is based on Gaussian mixture models that were trained and validated on full night acoustic signals that were recorded from a sleep laboratory, along with polysomnographic tests taken from patients with widely distributed severity of OSA. The snore detection system includes steps from noise reduction through event detection and all the way to snore identification.**

In order to analyze the performance of our proposed detector, a total of more than 80,000 acoustic episodes from 33 different OSA patients were manually segmented into snore and non-snore episodes; among the non-snore episodes we can find a variety of sleep related noises such as blanket and pillow murmurs, moaning, groaning, coughing, and talking. The validation dataset was recorded using two different audio recorders to ensure the robustness of the detector.

The events' total identification rate was 97.12% with 96.02% positive detection of snore as snore (sensitivity) and 97.90% detection of noise as noise (specificity).

Keywords: **Obstructive Sleep Apnea, Snore detection, GMM**

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a common sleep related breathing disorder in which the upper airways (UA) are collapsed, causing rapid and shallow breathing (hypopnea) or even total prevention of inhalation for at least 10 seconds (apnea), causing suffocation and frequent arousal during sleep. The main consequences of OSA are daytime sleepiness and increased risk of severe cardiovascular diseases, resulting in high risk of strokes and even sudden death [1,2].

Today, the gold standard for OSA diagnosis is polysomnography (PSG) [3] study, which requires a whole night diagnosis at a sleep laboratory while the subject is connected to numerous sensors; this study is expensive and the waiting list is long; Moreover, during this procedure sleep conditions are unnatural; these issues lead to seeking alternative methods of OSA diagnosis.

Snoring is the most common symptom of OSA, occurring in 70% to 95% of patients [4]. Snoring is caused by the vibration of soft tissues due to turbulent airflow through a narrow oropharynx in the UA [5], such a narrow oropharynx is more common among patients with OSA than subjects without OSA [6]. Earlier studies [7,8] suggested that the snores may play a key-role in detecting and distinguishing between healthy (non-OSA) and OSA patients. Since snores can be recorded using a non-contact microphone in any place, even at patients' homes, natural sleep can be obtained, and snore event detection can be used as the first stage of OSA detection system using audio signals.

In recent years, several snore/non-snore classification techniques have been published. Duckitt et al. [9] proposed a classification method for snore/non-snore episodes using mel-frequency cepstral coefficients (MFCC) with hidden Markov model (HMM) and achieved a detection rate of 82%-89% (from 6 simple snorer subjects). Cavusoglu et al. [10] proposed a method using sub-band spectral energy distributions along with robust linear regression (RLR) and principal component analysis (PCA); their detection rate was around 90.2% (using 15 subjects for each design and validation, ~9000 simple & OSA snore episodes in total).

We propose a Gaussian mixture model (GMM)–based method for snore/non-snore detection that involves acoustic feature extraction from three different feature-space domains: time, energy, and frequency. The proposed system produces a detection rate of 97.12% for snores and noise, using a real OSA population that was referred to PSG study. The system is robust for variety of snores, regardless of the subjects' gender or their OSA severity.

II. METHODS

Thirty-three patients (over 18 years old) scheduled for the sleep laboratory, were recorded during one night with a digital audio recorder device (EDIROL R-4) using a directional microphone (RODE NTG-1) at a distance of 1 meter above the head level and stored along with the PSG signals; the acquired audio signals are digitized at a

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

sampling frequency of 44.1 kHz, PCM, 16 bits per sample.

The raw audio signal is processed using the proposed snore detection system which is shown in Fig. 1.

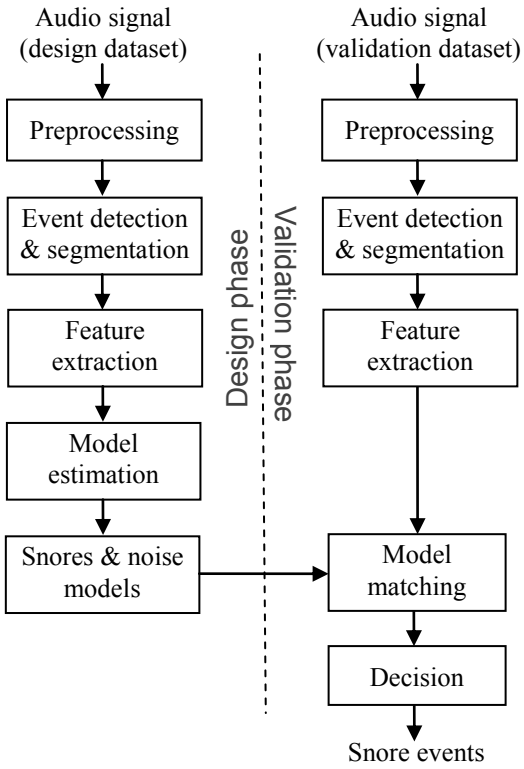


Fig. 1. – Block diagram of the snore detection system

Preprocessing

The signal is down-sampled to 16 kHz. A noise reduction (spectral subtraction) algorithm is applied on the full night audio based on the Wiener-filter, which is based on tracking a priori SNR using the decision-directed method proposed by Scalart et al. [11].

Event detection & segmentation

The audio events were automatically detected and segmented using an adaptive energy threshold.

In order to do so, the event detection module must include a few steps in order to achieve a potential snore event. At first the full night audio signal is analyzed in chucks of one-minute segments, for every segment an energy vector is calculated with a frame size of 60 ms and 75% overlap; the combined energy vectors are stored and will be used in the following steps. Later, a threshold is calculated using the estimated probability density function (pdf) of the energy values, where the first peak of the pdf is considered to be the background noise. With the completion of calculating every minute of the entire audio file, a smoothing technique based on median values is applied on the threshold vector to smooth outlier values.

With every energetic event that surpasses the relevant threshold, a boundary fine tuning technique is applied, based on the slopes of the linear regression fitting curves of 10 energetic samples (150 ms window) outside the event region on both sides; the event region is increased on each side as long as the slope does not change its sign.

Next is the fragmentation test – some of the events that are too close to each other (< 200 ms) are suspected to be a fragmented event (such as split snores); these events undergo a spectral similarity test of the 100 ms adjacent windows; in case of similarity the events are merged to form one event.

In order to improve detection efficiency, an event duration test is applied to remove unlikely snore events; only 200 ms to 3500 ms events are considered to be a potential snore event and sent to the snore classifier model.

Feature extraction

At this stage from each suspected event a 40-dimensional feature vector is calculated, consisting of three different sets of features:

1) Energy set

We included seven features such as skewness and kurtosis both for energy distribution in amplitude and in time, a normalized area beneath the energy envelope when a square shape represents one, a volume density rate [(max-min)/max] of the energy, and slope, which represents the slope from the beginning to the highest peak within the energy normalized duration.

2) Spectral-domain set

Twenty-seven features were included in this set. First we calculated 20 MFCCs for every 16 ms window (with a 50% overlapping) of the entire event; from that MFFC matrix, some of the features are extracted. We included the median (along time) of the first 16 of 20 coefficients from the MFCC matrix.

Two dynamic MFCC's distance (d_1 and d_2), which measure the MFCC's variance along time (1):

$$d_1 = \frac{1}{20}\sum_{k=1}^{20} VAR[MFCC(k,n)] \quad (1)$$

Where the variance (VAR) of MFCC(k,n) is estimated along time n and another version of distance uses the derivatives of MFCCs (2):

$$d_2 = \frac{1}{20}\sum_{k=1}^{20} VAR\left[\frac{d}{d\tau} MFCC(k,n)\right] \quad (2)$$

We also included a spectral flux, which was measured as the variance of the DFT along time (32 ms window duration, amplitude in dB).

A four sub-band frequencies distribution with a bandwidth of 2 kHz each – but only the first three was

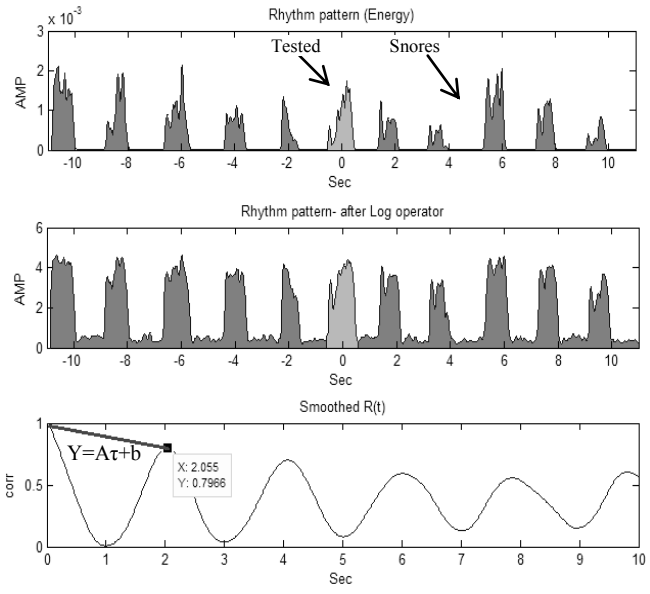
taken, frequency centroid, and the difference between the centroid of the initial half episode and the second half episode over time.

Pitch related features were added as well, such as pitch, pitch strength, and pitch density [8].

3) Time-domain set

Six features were included in this set such as episode duration, zero-crossing rate, rhythm period, and period strength; For the two last features we seek for a snoring pattern (evenly repeated events) which is calculated via autocorrelation of a 20 sec interval which includes the energy signal of the event surroundings; the rhythm period is the location (in time) of the first $R(\tau)$ peak. When the rhythm strength is measured as the product of the peak value of the first $R(\tau)$ and the variance between the $R(\tau)$ curve and the $A\tau+b$ linear regression fitting of $R(0)$ to that peak, the more "delta" shape the $R(\tau)$ the greater the error and therefore the greater the strength of the rhythm. A demonstration of snores rhythm is shown in Fig. 2; while this feature alone cannot be relied on, with the addition of the rest of the features, its contribution is a major addition in discrimination between snores and uncorrelated noises, but cannot be relied on alone.

We also added the ratio between forward and backward rhythm periods of the adjacent events as well as the strength of the ratio measured as the root square of the product between forward and backward strengths.



Demonstration of snores rhythm

Fig. 2. – The upper figure represents the energetic pattern of snores when the middle event is the tested snore. To emphasize the rhythm, a log operator and rescaling is applied to the energetic segment as shown in the middle figure. At the bottom is the autocorrelation R(t) of the segment.

Model estimation

In the design phase, two GMM-based models are estimated, one for snore events and one for non-snore events (model order 7 for snores and order 32 for non-snores) using feature vectors from the first 20 patients (design data). For this process, manually labeled events were used.

Model matching

Using the estimated models and the feature vectors, calculated from each event from the validation dataset (13 patients), a general classification decision (snore/non-snore) is performed using log-likelihood ratio (LLR) scores.

Decision

An adaptive LLR threshold is calculated using all the scores to assemble a pdf in order to find a minimum between the bi-modal Gaussian densities.

Among the events we recorded 35000+ noise events and 46000+ snores – both from simple and OSA snorers; the noise events were assembled from breathing, talking, blanket noises, and other non-snore events.

The tested group (for validation) contained 13 subjects, three of whom were recorded using another portable hand recorder (Olympus SL5, sampling frequency of 44.1 kHz, PCM, 16 bits per sample) located on the dresser beside the pillow in order to see if different recording devices and microphones can be used, although it was not included in the training (design) process.

III. RESULTS

The experiment was conducted using the database that is shown in Table 1.

Table 1 –Subjects' database information

	All	System Design	System Validation
# Subjects	33	20	13
Gender(M/F)	18/15	9/11	9/4
AGE - range	25-82	37-82	25-81
(mean ± std)	51.9±12.4	54.0±10.9	48.1±14.5
AHI - range	2.2-64.9	2.2-64.9	5.9-47.9
(mean ± std)	18.3±15.3	16.9±16.6	20.7±13.1
BMI - range	22.9-39.1	26.4-38	22.9-39.1
(mean ± std)	29.5±6.9	28.7±7.5	31.0±5.8
Snores (M/F)	23125/22109	16127/17502	6998/4607
Noise Events (M/F)	21486/13685	8457/9971	13029/3714
Device (RODE NTG-1/LS-5)	30/3	20/0	10/3

The design set for the model estimation has almost an equal number of men and women and a wide range of OSA severity – AHI of 2–65. For the validation set we included also three recordings from a handy recorder as shown in Table 1; we deliberately validated episodes recorded from a handy recorder in order to eliminate over fitting of the signals to our microphone's specs. The overall detection rate was 97.12% with 96.02% for snores and 97.90% for noise, with a confusion matrix as shown in Table 2.

Table 2 – Classification results

Class. As / True label	Snore	Noise
Snore	**96.02%**	3.98%
Noise	2.10%	**97.90%**

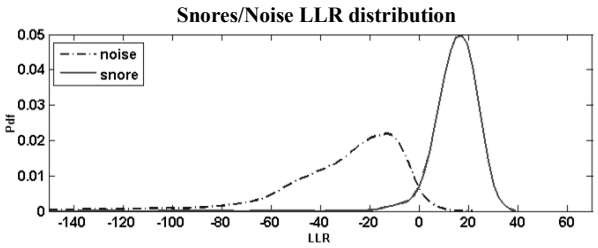


Fig. 3. – Log likelihood ratio (LLR) scores of snore and noise events

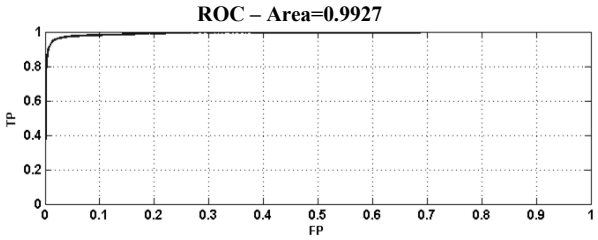


Fig. 4.– ROC curve – detection rate (True positive, TP) vs. (False Positive, FP) of snores

IV. DISCUSSION

The experiment was conducted on a total of 80,000+ audio events (snores and noise), which were extracted from 33 subjects as shown in Table 1, simple and OSA snorers, men and women – ensuring the robustness of the classifier. Furthermore, the validation group was assembled from signals that were recorded from different devices and even from different angles and distances, ensuring that the classification algorithm is robust to microphone type and angle; this implies that a small high-quality audio recording device can be used for home recordings, keeping the natural sleep of the patient.

According to the ROC curve in Fig. 4 and the LLR distribution in Fig. 3, we noticed that most of the errors were caused by a lack of distinction between smooth snores and breaths.

V. CONCLUSION

This paper proposed a snore detection system. The performance of the system is very encouraging – the detection rate is superior to earlier reported papers [9,10] and the system is ready for the next step – classification of OSA patients using snore analysis.

REFERENCES

[1] H.K. Yaggi, J. Concato, W.N. Kernan, J.H. Lichtman, L.M. Brass, V. Mohsenin, "Obstructive sleep apnea as a risk factor for stroke and death", N Engl J Med, vol. 353, pp. 2034-2041, 2005.

[2] A. Tarasiuk, S.G. Dotan, T. Simon, T. Tal, A. Oksenberg, H. Reuveni, "Low socioeconomic status is a risk factor for cardiovascular disease among adult obstructive sleep apnea patients requiring treatment", Chest, vol. 130, pp. 766-773, 2006.

[3] W. Flemons, "Obstructive sleep apnea", N Engl J Med, vol.347, pp. 498-504, 2002.

[4] M. Partinen, T. Telakivi, "Epidemiology of obstructive sleep apnea syndrome", Sleep, Vol.15, pp. S1-S4, 1992.

[5] V. Hoffstein, "Snoring", Chest, vol. 109, pp. 201-222, 1996.

[6] A. Malhotra, D.P. White, "Obstructive sleep apnea", Lancet, vol. 360, pp.237-245, 2002.

[7] A.K. Ng, T.S. Koh, E. Baey, T.H. Lee, U.R. Abeyratne, K. Puvanendran, "Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea?", Sleep Medicine, vol. 9, issue. 8, pp. 894-898, 2008.

[8] N. Ben-Israel, A. Tarasiuk, Y. Zigel, "Nocturnal sound analysis for the diagnosis of obstructive sleep apnea", Conf Proc IEEE Eng Med Biol Soc., pp. 6146-6149, 2010.

[9] W. Duckitt, S. Tuomi, T Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data", Physiol Meas, vol. 27, pp. 1047-1056, 2006.

[10] M. Cavusoglu, M. Kamasak, O. Erogul, T. Ciloglu, Y. Serinagaoglu, T. Akcam, "An efficient method for snore/nonsnore classification of sleep sounds", Physiol Meas, vol. 28, pp. 841-853, 2007.

[11] P. Scalart, J.V. Filho, "Speech enhancement based on a priori signal to noise estimation", Conf Proc IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp.629-632, 1996.

AN AUTOMATIC AND EFFICIENT METHOD OF SNORE EVENTS DETECTION FROM SLEEP AUDIO RECORDINGS

F. Gritti[1], L. Bocchi[1], I. Romagnoli[2], F. Gigliotti[2] and C. Manfredi[1]

[1] Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy
[2] Fondazione Don C. Gnocchi, Impruneta, Firenze, Italy

Abstract: **Snores are respiratory sounds produced during sleep. They are reported to be a risk factor for various sleep disorders, such as obstructive sleep apnea syndrome (OSA). Diagnosis of OSA relies on the expertise of the clinician that inspects whole night polysomnographic recording. This inspection is time consuming and uncomfortable for the patients. Thus, there is a strong need for a tool to analyze snore sounds automatically. Nocturnal respiratory sounds are composed of two kind of events: "silence" episodes and "sound" episodes that include breathing, snoring and "other" sounds.**

In this paper a new method to detect snoring episodes from full night audio recordings is proposed. Signal analysis is performed in three steps: pre-processing, automatic segmentation, extraction of features and classification. With the segmentation step, only the "sound" parts of the audio signal are extracted using a Short-Term Energy and the Otsu thresholding method. The aim of classification step is the detection of snore episodes only, using two Neural Artificial Network applied to four features (length, maximum amplitude, standard deviation and energy).

Data from 24 subject are analyzed using the proposed method; on the dataset, a sensitivity of 86,2% and specificity of 86,3% are obtained.

Keyword: **Snore, Obstructive sleep apnea, Neural network, Automatic segmentation**

I. INTRODUCTION

Snoring can be defined as a respiratory noise that is generated during sleep when breathing is obstructed by a collapse in the upper air way. Loud and regular snoring is the earliest and most consistent sign of upper airway (UA) dysfunction leading to sleep apnea/hypopnea syndrome [1].

Obstructive sleep apnea (OSA) is the most frequent encountered form of the sleep apnea [1]. In OSA, the upper airways are obstructed during sleep, resulting in the decrease of oxygen flow to the lungs. Patients suffering from OSA often wake up frequently. When there is a full closure of airways, the disease is termed "apnea" while when there is a partial closure, it is known as "hypopnea" [2]. The disease is associated with significant clinical consequences but it is frequently unrecognized and undiagnosed because simple, low-cost devices for mass screening of the population do not yet exist.

The current "gold standard" method for sleep apnea assessment is Polysomnography (PSG). This technique requires a full night hospital during which the patient is connected to more than ten channels of measurements requiring physical contact with sensors. PSG is thus inconvenient, expensive and unsuited for community screening [3] [4] [5]. Thus, in order to study OSA non-invasively, several researches focused on the analysis of snore sounds from full night audio signal recordings, using signal processing techniques.

Commonly tracheal respiratory sounds are recorded using a microphone placed over the patient's neck or hung above the patient's head during the night, leading to long lasting audio signals (6–8 hours). The length of a whole recordings is thus prohibitive for the analysis by listening to and for visual inspection of signal patterns. Hence, automatic methods are needed to speed up the analysis task.

Despite its clinical relevance, a limited number of studies on automatic detection and classification of snore sound has been developed to date [6], [7], [8], [9], [10], [11]. In these works different kind of techniques of analysis are applied, such as: Energy and zero-crossing rate [6][7] [8], Hidden Markov Models (HMMs) and spectral-based features [9], 500Hz sub-band energy distribution [8],[10], normalized autocorrelation coefficient at 1 ms delay and the first predictor coefficient of LPC analysis [6], and frequency range of each formant [11].

However, most often the automatic segmentation step is not included, the snore events being detected manually or with semi-automatic methods.

Hence the motivation of this study was to develop an effective method to detect the snoring episodes, fully automatic and fast enough to allow processing full night recordings in a reasonable amount of time.

A short-term energy measure was implemented for automatic detection of "sound" events and two neural artificial network were applied to four features (length, maximum amplitude, standard deviation and energy), for automatic classification of snore events.

II. METHODS

The aim of the proposed system of analysis is the detection of snoring events from full night audio

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

recordings. This is achieved by means of the following three steps:

A. *Pre-processing*: loading of audio signal, band-pass filtering and down sampling;
B. *Automatic segmentation*: detection of the "sound" parts of the signal;
C. *Extraction of the features and classification*: identification of snoring events.

The implemented method, named Snore Analyzer, is developed under Matlab 7.11.00 software tool. A flow chart is shown in Figure 1.

Snore Analyzer is provided with a user-friendly interface (Figure 2) that easily allow the user to choose the audio signal to be processed (*Load* bottom) and set the following parameters for subsequent processing: 1) Sampling frequency (44.100 kHz by default); 2) Down sampling frequency (11.025 kHz by default); 3) Starting and ending samples, to select the part of the signal to be processed; 4) Size of analysis window (40 ms by default).

Then the user starts the elaboration of the selected audio signal pushing the *Start* bottom. Through the *Reset* bottom, the user can delete all the items.

The elaboration of whole signal (or a part of it) is fully automatic and the user should not act manually anymore.

The length of each audio signal is about 7-8 hours and the complete analysis of whole signal requires about 30-40 minutes. At the end, the software gives as output a list of extracted "sound" events which are labeled as snore or not-snore.

The next sections (A, B, C) describe each step in detail.

A. Pre-processing

The use of a robust recording system can improve signal acquisition, but noise reduction is required to eliminate interferences. Therefore a pre-processing step is implemented to improve signal to noise ratio.

In this study the audio signal is bandpass filtered by a Butterworth filter of order 5 and a cut-off frequency of 100 – 1000 Hz, to reduce the effects of heart sounds and high-frequency noises [1]. Main frequency components of breathing and snoring sounds are in fact included in this range [12] [13]. After the filtering step, the signal is down sampled (to 11.025 kHz), to reduce the size of the data and hence speed up signal processing.

B. Automatic Segmentation

The audio signal is typically a mixture of two different kind of events: "silence" that do not contain any sound and "sound" that include breathing episodes, snoring episodes and "other" sounds such as oral noise, ambient sounds, patient's cough, speech and blanket movements, etc.

This step is therefore devoted to identify the "sound" events. Short-Term Energy (STE) is a commonly used measure for determining the "sound" parts as it increases during "sound" events and decreases during "silence" episodes [14] [15].

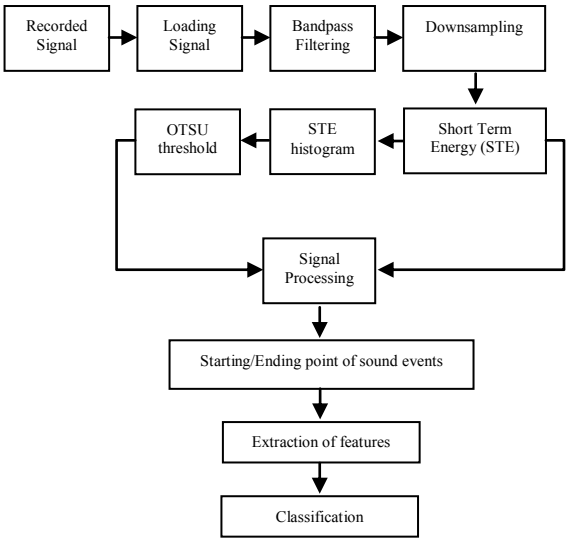


Fig. 1 Flow chart of the analysis system.

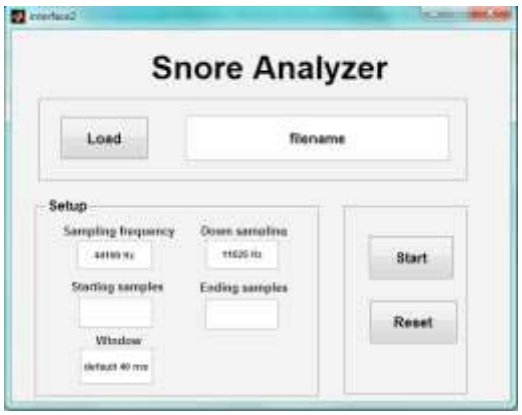


Fig. 2 User-friendly interface of the implemented software tool.

In our study, STE is evaluated in signal windows of 40 ms in length with 50% overlap between adjacent windows. In order to determinate boundaries of "sound" events, we computed the histogram of the signal energy and the Otsu method is iteratively applied to obtain two thresholds: the upper one t_u and the lower one t_l [16], [17]. These thresholds are then used to find the starting and ending points of each "sound" event in the audio signal. In particular, when the STE curve overpasses the upper threshold, the first point under the lower threshold (on the left side of the curve with respect to the upper threshold) is detected in order to get the starting point. When the STE curve falls down t_l, the ending point of the event is found (Figure 3).

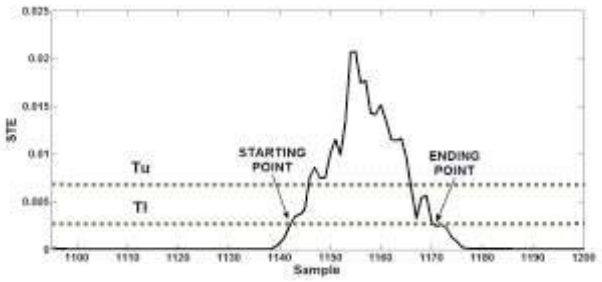


Fig. 3 The starting and ending points of a "sound" event.

At the end of segmentation, the list of extracted "sound" events is saved in a text file to be used in the classification step.

C. Extraction of the features and Classification

Once all the "sound" events from the signal are obtained, they have to be classified as snore or not-snore (i.e. breath and "other" events). In fact for a reliable analysis of OSA, only snore episodes must be detected. This task is carried out in two step: in the first on, a set of four parameters is computed in time domain; in the second one, the events are identified with a classification system.

The first parameter is the length of each "sound" event, calculated as the distance between the starting and the ending point of the event. This feature allows to distinguish between "other" events and breathing/snoring, as the average length (in samples), computed for breathing and snoring sounds, is lower than for "other" sounds, as shown in Table 1.

Table 1 Mean and Standard Deviation of the length of snore/breath and of "other" sounds.

Length [sample]		
	Mean Value	**STD Value**
Snore/Breath	$4.7999 \cdot 10^4$	$2.8492 \cdot 10^4$
"Other"	$1.4405 \cdot 10^4$	$1.3602 \cdot 10^4$

The other parameters are: the Standard Deviation (STD), the mean value of Short-Term Energy (STE) and the maximum amplitude of "sound" events, given by the difference between the maximum and the minimum amplitude of the signal.

These features allow to distinguish between snoring episodes and breathing episodes, as the average value of each single feature is higher in the class of snoring events than in the class of breathing events (Table 2) while the behaviour of these parameters is highly variable in "other" sounds.

Hence the following observation can be made: "other" sounds can be found using the length of the events only; snoring and breathing sounds can be distinguished using the STD, the mean value of the STE and the maximum amplitude.

Table 2 Mean and Standard Deviation of STD, mean of STE and Maximum Amplitude.

STD		
	Mean Value	**STD Value**
Snore	0.0038	0.0024
Breath	0.0014	0.0005

STE		
	Mean Value	**STD Value**
Snore	-5.4120	0.5283
Breath	-6.0974	0.3061

Maximum Amplitude		
	Mean Value	**STD Value**
Snore	0.0498	0.0389
Breath	0.0142	0.0061

According to these results, a classifier is designed made up by two artificial neural networks: the first one is used to identify the "other" sounds, while the second one is used to discriminate between snoring and breathing sounds.

The sounds episodes were manually labelled by trained clinicians as snore or not snore to built the training and the testing datasets for the classification system. The training set is made up by 1643 sound signals equally distributed among snoring, breathing and 'other' sounds.

The first network is trained with all the events of the training set using only the length of the event as input and the outcome of listening is used as teaching input.

After the training step, the network output is tested and compared with the outcome of listening; the "other" sounds correctly recognized as "other" (true negative) are removed from the training set used in the second network that consists of three inputs, corresponding to the mean value of STE, its STD and the maximum amplitude, respectively.

III. Results

Clinical audio signals (18 patients of different age and sex) are recorded at Fondazione Don Gnocchi, Pozzolatico, Firenze, where the patients slept in single bedroom, separated from television and others predictable sources of noise.

The audio signal are digitized at 16-bit with a sampling frequency Fs=44.100 kHz, using a Tascam Us-144 sound card and a unidirectional microphone Shure SM58, positioned at about 30 cm from the mouth of the

patient. The length of single signal is of about 7-8 hours, but, for the analysis, we considered thirty minutes of each recording, selected in the central part of the signal when the patient was sleeping and low environmental noise was detected.

A preliminary evaluation was carried out to assess the performance of the automatic segmentation, evaluated as the percentage of sounds detected over the total number of sounds, Resulting in about 97%.

Concerning the classification step, the first network was tested on 787 "sound" events, different from the original training set. From the analysis of the ROC curve, a "best" threshold was obtained that allows to correctly identify 85.4% of the "other" sounds. These sounds were stored in a list of not-snore events and removed from the test set.

The second network was tested on the remaining sounds and, as for the first network, the best ROC threshold was computed and used to identify snore and not-snore sounds.

The accuracy (number of correct classifications) of the second network was found equal to 86.2%. This result corresponds to a sensitivity (true positive (TP) ratio) equal to 86.2 and a specificity (true negative (TN) ratio) equal to 86.3.

IV. DISCUSSION AND CONCLUSIONS

A full automatic and unsupervised system for snore identification during sleep is proposed.

The proposed automatic segmentation was shown to be a reliable technique for the extraction of sound events as almost all silence events were discarded.

The algorithm for classification correctly identifies the 86.2% of analysed events. However it fails in case of low intensity snores, as such events have low energy and low maximum amplitude. But, as post apnoeic snore events are usually more intense than non-post apnoeic ones, this limitation could be acceptable.

Future work will be devoted to enhancing the procedure, increasing the dataset and defining a reliable method for the identification of post-apnoeic events from the automatically detected snore sounds, e.g. as in [18].

REFERENCES

[1] M.J. Thorpy, "*The international classification of sleep disorders: diagnostic and coding manual*", Lawrence KS, ed. Allen Press Inc., USA, 1990, pp. 195–197.

[2] N.J. Douglas, *Harrison's Principles of Internal Medicine*, McGraw-Hill , 17th ed. New York, 2008.

[3] A. Ayatollah. and Z. Moussavi, "Automatic breath and snore sounds classification from tracheal and ambient sounds recordings", *Medical Engineering & Physics*, vol. 32, pp. 985-990, 2010.

[4] W.W. Flemons, M.R. Littner, J.A. Rowley, P. Gay, W.M. Anderson, D.W. Hudgel, R.D. McEvoy, and D.I. Loube, "Home diagnosis of sleep apnea: A systematic review of the literature," *Chest*, vol. 124, pp. 1543–1579, 2003.

[5] C.A. Kushida, M.R. Littner and T.Morgenthaler, "Practice parameters for the indications for polysomnography and related procedures: An update for 2005," *Sleep*, vol. 28, no. 4, pp. 499–521, 2005.

[6] A.S. Karunajeewa, R. Abeyratne and C. Hukins, "Silence–breathing–snore classification from snore-related sounds", *Physiol. Meas.*, vol. 29, pp. 227–243, 2008.

[7] U.R. Abeyratne, A. Wakwella and C. Hukins, "Pitch jump probability measures for the analysis of snoring sounds in apnea", *Physiol. Meas*., vol. 26, pp. 779–798, 2005.

[8] A. Ayatollah. and Z. Moussavi, "Automatic and Unsupervised Snore Sound Extraction From Respiratory Sound Signals", *IEEE Transaction on Biomedical Engineering*, vol. 58, n.5, pp. 1156-1162, May 2011

[9] W.D. Duckitt, S.K. Tuomi and T.R. Niesler, "Automatic Detection, segmentation and assessment of snoring from ambient acoustic data", *Physiol. Meas.*, vol. 27, pp. 1047-1056, 2006.

[10] M. Cavusoglu, M. Kamasaka and O. Erogul, "An efficient method for snore/notsnore classification of sleep sounds", *Physiol. Meas*., vol. 28, pp. 1-13, 2007.

[11] A. Yadollahi and Z. Moussavi, "Formant analysis of breath and snore sounds", *Proc. IEEE EMBS, Minneapolis*, MN, 3-6 Sept. 2009, pp. 7110 – 7113, 2009.

[12] J. Fiz, J. Abad, R. Jane, M. Riera, M.A. Mananas and P. Caminal, "Acoustic analysis of snoring in patients with simple snoring and obstructive sleep apnea", *Eur Respir J*, vol. 9, pp. 146-59, 1996.

[13] R. Beck, M. Odeh, A. Oliven and N. Gavriely, "The acoustic properties of snores", *Eur Respir J*, vol. 8(12), pp. 2120-8, 1995.

[14] J R. Deller, J.H. Hansen and J.G. Proakis, *Discrete-Time processing of speech signal*., Mcmillan Publishing Company, New York, 1993, pp. 724-728.

[15] A. Kulkas, E. Huupponen, "Intelligent methods for identifying respiratory cycle phases from tracheal sound signal during sleep", *Comput Biol Med*, vol. 39, pp. 1000-1005, 2009.

[16] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Sys., Man., Cyber*, vol. 9(1), pp.62-66, 1979

[17] M. Calisti, L. Bocchi, C. Manfredi, I. Romagnoli, F. Gigliotti and G. Donzelli, "Automatic detection of post-apnoeic snore events from home and clinical full night sleep recording, *MAVEBA 09*, Florence, Italy, pp.189-192, 2009

[18] E. Goldshtein, A. Tarasiuk, and Y. Zigel, "Automatic Detection of Obstructive Sleep Apnea Using Speech Signals" *IEEE Transaction on Biomedical Engineering*, vol. 58, n.5, pp. 1373-1382, May 2011

Session II:
Imaging

ELECTRICAL IMPEDANCE TOMOGRAPHY IMAGING OF LARYNX

A. Seppänen[1], A. Nissinen[1,2], V. Kolehmainen[1], S. Siltanen[3], A-M Laukkanen[2]

[1]Department of Applied Physics, University of Eastern Finland, Kuopio, Finland
[2] Speech and Voice Research Laboratory, School of Education, University of Tampere, Finland
[3]Department of Mathematics and Statistics, University of Helsinki, Finland

Abstract: **In this paper we discuss electrical impedance tomography (EIT) imaging of human larynx. Especially, we focus on monitoring of vocal folds. EIT is a non-invasive three-dimensional (3D) imaging modality based electrical measurements conducted from the skin of a person. We hypothesize that EIT reconstructions can provide information on vocal folds' movement as well as physiological changes in vocal fold tissue caused by the vocal loading. This information could be used for quantifying vocal loading and measuring the consequences of vocal loading. In this paper, the feasibility of EIT for imaging of larynx is tested with numerical simulation studies. The preliminary results suggest that EIT is sensitive to movement of vocal folds.**

Keywords: **Electrical impedance tomography (EIT), imaging, larynx, vocal loading**

I. INTRODUCTION

Electrical impedance tomography (EIT) [1] is an imaging modality based on non-invasive electrical measurements. In biomedical applications of EIT, an array of electrodes is attached on the skin of a person. Weak alternating currents are injected through chosen electrodes and the resulting potentials are measured on several electrodes. This procedure is repeated using various current injection patterns. Based on the collected current and potential data, the internal three-dimensional (3D) conductivity distribution is reconstructed. The biomedical applications of EIT include e.g. monitoring of ventilation and diagnosing breast cancer.

With the aid of EIT, it might also be possible to get information on human larynx: EIT could perhaps serve as a tool for imaging the vocal fold's movement during speech production, and for estimating the physiological changes in the vocal fold tissue caused by vocal loading. This information could be utilized for detecting and quantifying vocal loading (i.e. getting estimates of stresses acting upon the tissue) and measuring the consequences of vocal loading (i.e. changes in the tissue). Indeed, EIT has a high potential for glottal diagnostics. Basically the data used in EIT consists of measurements also used in electroglottography (EGG) [2,3], which is a regularly used tool in the assessment of voice production. However, while in standard EGG two-channel impedance measurement data is considered, EIT is based on multi-channel data. Moreover, EIT utilizes advanced mathematical modeling in the data processing. This enables estimation of spatial properties of the larynx, in addition to temporal change information provided by EGG. A dual-channel EGG has been used since the 1990's [3]. Recently, Kob and Frauenrath [4] proposed a multi-channel-EGG system for improving the assessment of glottal opening and the laryngeal position. The measurement setup was similar to EIT, but the data was not used for 3D image reconstruction. However, the results indicated that it is possible to track the location of glottis during a swallowing manoeuvre. In this paper, the feasibility of EIT for imaging of the larynx is discussed. Especially, the computational challenges associated with the complex internal structure of larynx are considered.

II. METHODOS

Mathematically, the image reconstruction problem in EIT – determining the conductivity distribution on the basis of the measured electrode potentials – is an ill-posed inverse problem. By definition, ill-posed problems are extremely sensitive to measurement noise and modeling errors [5]. In consequence, the accurate mathematical modeling of the measurements plays a paramount role in the successful EIT reconstruction. In addition, prior information on the conductivity needs to be incorporated in the reconstruction. The most accurate mathematical model for EIT measurements is the complete electrode model (CEM) [6] which consists of an elliptic partial differential equation and associated boundary conditions. The 3D finite element approximation of the CEM was presented in [7]. In this paper, the feasibility of EIT for imaging of the larynx is studied by a numerical simulation. The computational model is an adaptation of the model presented in [7]. The image reconstruction problem is formulated as a Bayesian inverse problem [5].

III. RESULTS

Figs. 1 and 2 show examples of modeling a larynx. In this 2D simulation study, the model was constructed based on a cross sectional image data obtained from computerized tomography (CT) of the larynx. The images

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

in Figs. 1 and 2 represent cases of a closed and an open glottis. In both cases, the EIT measurements were simulated by numerical approximation of the CEM. The placement of EIT electrodes is shown in the figures. Note that electrodes were set only in front of the neck. This choice was made because electrical measurements from back of the neck would not be sensitive to the conductivity of the glottis, due to the electrically insulating spine (whitish structure in Figs. 1 and 2). For a realistic simulation, Gaussian noise was added to the synthetic measurements.

The EIT reconstructions were computed assuming that the outer shape of the neck is known; the observation model for the inverse problem was constructed using this geometry. However, no information on the internal structures shown in Figs. 1 and 2 was utilized in the reconstructions. By contrast, a standard smoothness prior model [5] for the conductivity distribution was written. This model is clearly not well justified in the present case – the true conductivity distributions (Figs. 1 and 2) are not spatially smooth, because of the anatomical structures inside the larynx.

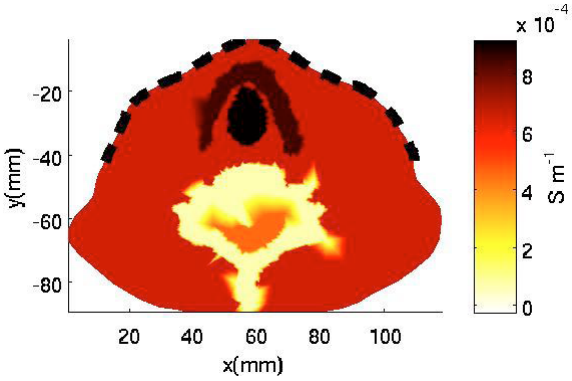


Figure 1. A conductivity distribution used in the simulation: glottis closed. The black bars in front of the neck represent the EIT-electrodes.

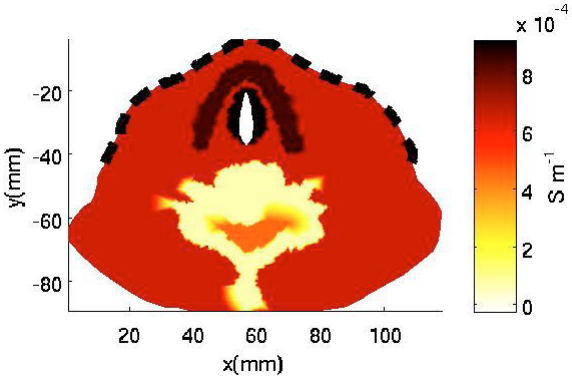


Figure 2. A conductivity distribution used in the simulation: glottis open.

The EIT reconstructions computed on the basis of the noisy observations are depicted in Figs. 3 and 4. In the case of a closed glottis (Fig. 3), the EIT reconstruction shows a blurry conductive region roughly covering the locations of the glottis and the surrounding cartilage. The internal structures cannot be detected accurately because of the relatively low spatial resolution of EIT, and especially because of the unsuitable prior model used in the reconstructions: the effect of the smoothness assumption is clearly visible in the reconstruction. In the case of an open glottis, the conductivity was reconstructed by taking the so-called difference imaging approach [8]. That is, the data of both measurement sets (corresponding to the open and closed glottis) were utilized by taking into account that the difference in the measured data is due to a change of the conductivity distribution. The reconstruction is shown in Fig. 4. The inclusion of low conductivity in the image is due to opened glottis; the opening acts as a perfectly insulating object. This inclusion is relatively well localized. Hence, although the resolution of individual images is quite poor, the difference in the state of the glottis can clearly be distinguished by comparing the two EIT reconstructions.

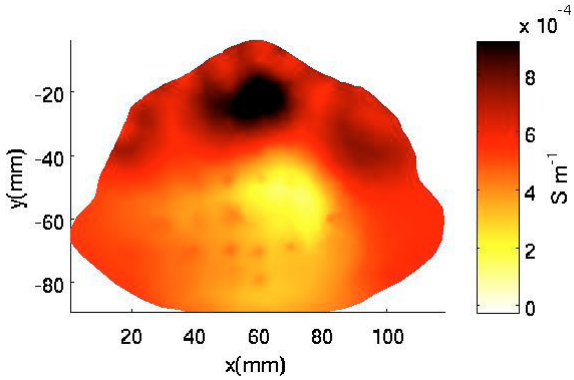


Figure 3. Reconstructed conductivity distribution in the case of a closed glottis.

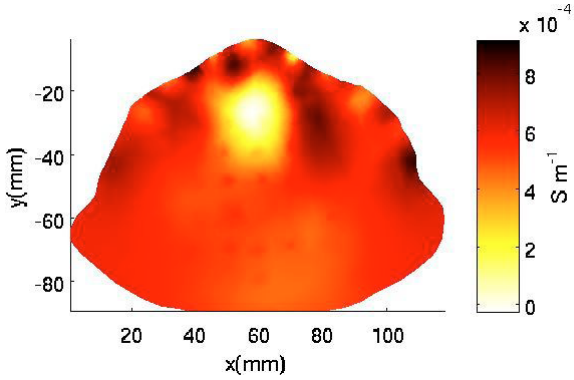


Figure 4. Reconstructed conductivity distribution in the case of an open glottis.

IV. Discussion

In the preliminary tests presented in the previous section, the aim was to study the feasibility of EIT for imaging the human larynx. Especially, we concentrated on the effect of complex internal structures on EIT reconstructions. The presence of these structures makes the problem of imaging the larynx more challenging than targets that exhibit more smoothness. A pleasant property of the larynx imaging problem, however, is the possibility of carrying out the reference measurements for difference imaging. Indeed, a test person can deliberately close or open the glottis for the reference measurements. The reference data is not available in many other biomedical applications.

V. Conclusion

In this paper, the feasibility of EIT to imaging the larynx was studied with 2D numerical simulations. The preliminary results suggest that by taking the difference imaging approach it is possible to get information on properties of glottis with EIT. Future research topics include 3D modeling of the larynx and modeling of uncertainties related to neck shape. Further, experimental studies will be carried out.

References

[1] G.J. Saulnier, R.S. Blue, J.C. Newell, D. Isaacson and P.M. Edic, "Electrical Impedance Tomography", *IEEE Signal Processing Magazine*, vol. 18, pp. 31-43, 2001.

[2] A. Fourcin and E. Abberton, "First applications of a new laryngograph", *Med. Biol. Illus.* vol. 21, pp. 172–182, 1971.

[3] M Rothenberg, "A multichannel electroglottograph", *Journal of Voice*, vol. 6, pp. 36-43, 1992.

[4] M. Kob and T. Frauenrath, "A system for parallel measurement of glottis opening and larynx position", *Biomedical Signal Processing and Control*, vol. 4, pp. 221–228, 2009.

[5] J.P. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*. Springer-Verlag, New York, 2005.

[6] K.-S Cheng, D. Isaacson and J.C. Newell, "Electrode models for electric current computed tomography", *IEEE Trans Biomed Eng*, vol. 36, pp. 918-924, 1989.

[7] P.J. Vauhkonen, M. Vauhkonen, T. Savolainen and J.P. Kaipio, "Three-dimensional electrical impedance tomography based on the complete electrode model", *IEEE Transactions in Biomedical Engineering*, vol. 46, pp. 1150 – 1160, 1999.

[8] B.H Brown, "Electrical impedance tomography (EIT): a review", *Journal of Medical Engineering & Technology*, vol. 27, pp. 97 – 108, 2003.

PVG-WAVEGRAMS: THREE-DIMENSIONAL VISUALIZATION OF VOCAL FOLD DYNAMICS

J. Unger[1], T. Meyer[1], C. T. Herbst[2], M. Döllinger[3], J. Lohscheller[1]

[1] University of Applied Sciences Trier, Department of Computer Science, Trier, Germany
[2] University of Vienna, Department of Cognitive Biology, Vienna, Austria
[3] University Hospital Erlangen, Department of Phoniatrics and Pediatric Audiology, Erlangen, Germany

Abstract: **Recently, endoscopic high-speed laryngoscopy has been established for commercial use and constitutes a state-of-the-art technique to examine vocal fold dynamics.**
Due to the need of high sampling rates, a high amount of frames has to be considered for subjective assessment. Especially long phonation recordings produce several hundred megabytes of digital data. We present a technique for visualizing these high-speed videos in a compact and intuitive form. The high-speed videos are therefore mapped to three-dimensional cycle-based graphs representing a detailed visualization of vocal fold dynamics.
Keywords : **High-speed-laryngoscopy, visualization, vocal fold dynamics, voice assessment**

I. INTRODUCTION

Investigation of vocal fold (VF) dynamics is not only essential for understanding the mechanism of voice production. It plays also an important role in voice assessment and treatment of voice disorders.

Recently, endoscopic high-speed laryngoscopy has been established for commercial use and constitutes a state-of-the-art technique for analyzing VF vibratory behavior in vivo. Modern cameras provide sampling rates of usually 2,000 to 6,000 frames per second. Hence, huge amounts of data have to be considered for visual assessment and analyzing purposes. Especially investigations of long and non-stationary phonation sequences remain laborious and time consuming.

To avoid visual inspection of motional processes of video sequences, compact VF visualization and analysis has been addressed by several authors: The Hilbert transformed of the glottal area waveform is utilized by Yan et al. [1]. The Nyquist plot and the envelope characterize the obtained analytic signal in terms of perturbation and periodicity. However, glottal area does not differentiate between ventral and dorsal as well as left and right VF oscillations. Li et al. [2] obtain spatial and temporal eigenfolds through singular value decomposition of vocal fold movement at different locations along the glottal axis. The first, second and third eigenfold reflect the average shape, the closing pattern of the VFs and the motion of the VFs in longitudinal direction. The separation of spatial and temporal eigenfolds offers a compact visualization but does not allow to localize vibratory features in space and time simultaneously. Therefore, a mapping of laryngeal high-speed videos into two dimensional color graphs is brought up by Lohscheller et al. [3]. Phonovibrograms (PVGs) and Glottovibrograms (GVGs) give a detailed representation of laryngoscopic high-speed videos but are mainly designed to characterize healthy and abnormal VF vibratory patterns .

Herbst et al. [4] developed a technique for visualizing electroglottographic signals in compact cycle-based graphs called wavegrams. Electroglottography provides a relative measure of vocal fold closure by quantifying the impedance of an alternating voltage between two electrodes placed on the surface of the throat at the level of the thyroid cartilage. Electroglottographic impedance measures are therefore restricted to general VF contact phenomena and provide only rough information of VF dynamics.

Currently, there is still a need of visualization methods which allow detailed representation of non-stationary and long phonation sequences. This paper presents a visualization method of laryngoscopic high-speed videos combining and extending the ideas of PVGs, GVGs and electroglottographic wavegrams, reflecting VF dynamics in three-dimensional graphs.

II. METHODS

The construction process of PVG- and GVG-wavegrams comprises data preprocessing and graph assembling steps. The preprocessing part extracts PVGs, GVGs and glottal area from high-speed videos. For a detailed description of glottal area segmentation, PVG- and GVG assembling see Lohscheller et al. ([5], [3]).

The wavegram assembling process transforms PVG and GVG data to three-dimensional cycle-based graphs. Therefore, a cycle detection routine is performed on glottal area signal. Cycles are separated by determining points in time t_i corresponding to the i^{th} maximum of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

glottal opening. For this purpose, auto-correlation of the first 100 frames estimates the fundamental frequency *f* at the beginning of the signal. The maximum value within a window with a length of $T=\frac{1}{f}$ yields t_1 corresponding to the first maximum glottal opening. If this peak is located in the first half of the window, it is shifted by $\frac{3}{4}T$ in positive time direction, in the other case it is shifted by $\frac{5}{4}T$. This procedure is repeated until the window reaches the end of the signal. The i^{th} cycle is then separated by cutting the interval

$$[t_i-\frac{t_i-t_{i-1}}{2}, t_i+\frac{t_{i+1}-t_i}{2}] \quad (1)$$

The first and the last cycle may have been truncated, therefore, they are being disregarded in the following analysis.

The separated cycles are concatenated along a new axis in such a manner that one axis addresses different cycles and the other displays the normalized cycle progress. The individual cycle length l_i is normalized and associated with the cycle width in time direction (see Fig. 1)

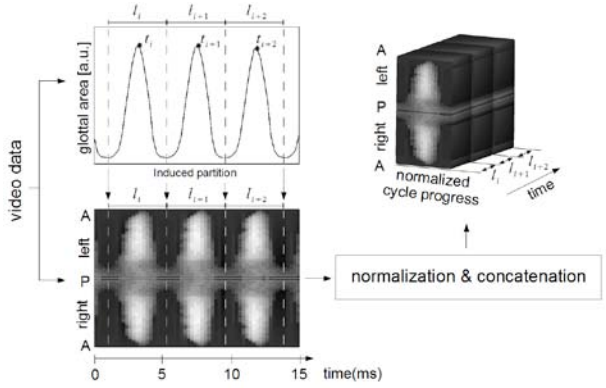


Figure 1: Glottal area and PVG are extracted from video data. The glottal area induces a cyclewise partition of PVG data (left side). The concatenation process increases the dimensionality resulting in a three-dimensional PVG graph.

Visualization of three dimensional PVG and GVG data requires rendering techniques, projecting three dimensional graphs on two dimensional monitor screens. A great variety of rendering techniques established in computer graphics. Among these, we will depict two different methods for our purposes:

Isocontour extraction [6] reduces three dimensional data to contours with constant intensity. Thus, a single parameter suffices to define an isocontour surface. For visualizing PVG data, isocontour surfaces illustrate points with constant deflection ranging from 0 to 100 percent of maximum deflection (Fig. 2, left side).

Volume Rendering Techniques [7] utilize transfer-functions for mapping each volume element (voxel) from 3D data to opacity and color. Thereby, the transfer function specializes the region of interest in terms of high opacity values and allows visualization of interior regions if opacity is low. For our purposes, we defined a transfer function with linear ascend in color intensity (dark colors correspond to slight deflections) and opacity is mapped to the normalized absolute value of the deflection (Fig. 2, right side).

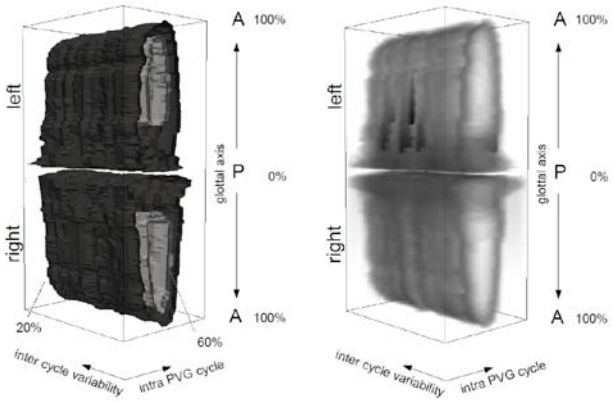


Figure 2: Visualization techniques for PVG data computed from 0.125 seconds of sustained stable phonation: (a) Isocontour extraction and (b) Volume Rendering

To demonstrate our visualization method, one male subject (30 yrs, no known voice disorders), was introduced to (a) maintain a stable phonation at habitual loudness and pitch; and (b) to produce a glissando from modal to falsetto register during endoscopic examination. The laryngoscopic video was recorded using the Endocam 5562 high-speed camera system (Richard-Wolf GmbH, Knittlingen, Germany).

III. RESULTS

The fundamental frequency during the glissando is depicted in Fig. 3. Four labels are marked on the time axis: The first (beginning until A) and the last part (D until end) have nearly constant frequency of 140Hz and 390Hz, respectively. Within the PVG- and GVG-wavegrams, characteristic changes of the vibratory pattern of vocal folds can be identified in terms of geometric shapes and colors. Figure 4 shows the GVG-wavegram of the non-stationary glissando. The GVG-wavegram revealed significant changes in vibratory patterns during pitch-raise between A and D: At the beginning the glottis opened alongside the entire vocal fold length l_1 and at the end vocal folds opened merely within anterior parts along l_2 .

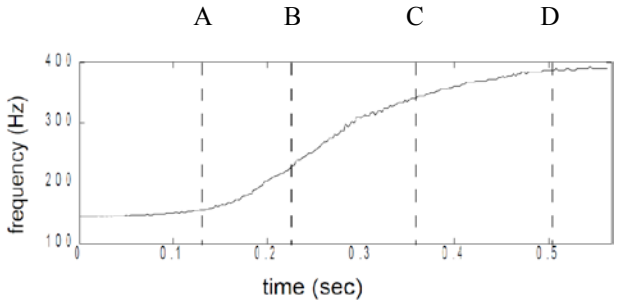


Figure 3: Change of fundamental frequency during pitch-raise

The position of the oscillation amplitude along the glottal axis, emphasized by the dashed line, moved towards anterior. The projection along the glottal axis (Fig.4, bottom) reveals the glottal area represented as 2D-wavegram. Two features were obtained from 2D wavegram representation: First, the transition from dark to light color intensities in time direction implied a decreasing glottal area and secondly, dark and white

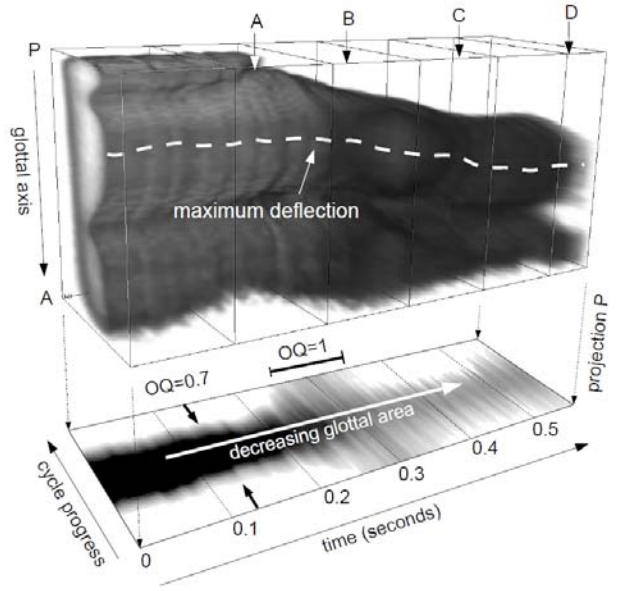


Figure 4: GVG wavegram of a pitch raise from 140 to 390 Hz. The projection along the glottal axis, represented as 2D-wavegram, provides information of the glottal area.

regions constituted opened and closed states of the glottis and allows determination of e.g. the open quotient (OQ).

The PVG-wavegram allows a separated analysis of left and right vocal fold. It revealed symmetric vibratory patterns of left and right vocal fold (Fig. 5). The dashed lines mark the oscillation amplitude positions over time separated for left and right vocal fold.

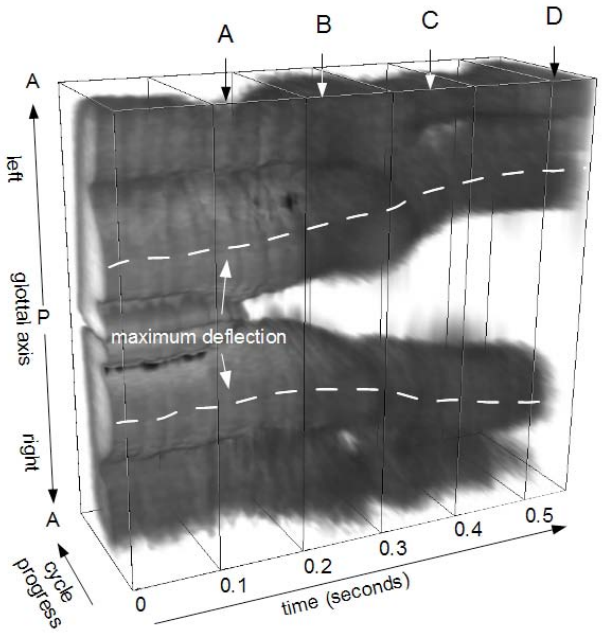


Figure 5: The PVG-wavegram allows the distinction of left and right vocal fold oscillations

IV. Discussion

For an detailed and compact visualization of vocal fold dynamics, three dimensional graphs have been developed combining and extending the ideas of detailed and intuitive PVG- and GVG-representation with the compact cycle-based wavegram-technique. PVG- and GVG-wavegrams provide a new and powerful method for visualizing vocal fold dynamics of especially long phonation sequences in a single graph. We showed that information of oscillating amplitudes and their position along the glottal axis can be obtained from PVG- and GVG-wavegrams.

Opened and closed states and clinically established features like open- and speed quotient as well as glottal amplitudes are embodied in the corresponding projection along the glottal axis.

PVG-wavegrams enable a separate visualization of left and right vocal fold vibration: Vibratory patterns, asymmetries as well as amplitudes of left and right vocal fold oscillation are emphasized by PVG-wavegrams.

In addition, characteristic geometric PVG-features, which have been already used for classifying pathological vocal fold vibrations [8], are represented intuitively.

V. Conclusion

PVG- and GVG-wavegrams aim to visualize vocal fold dynamics obtained from segmented laryngoscopic high speed videos. They accent the shape of the glottal cycles and its time development and therefore, they are useful

for classification of pathologies. Approved features like amplitudes, asymmetries, open- and speed quotient are embodied in geometrical shapes. Further investigations of characteristic geometric shapes will help to interpret PVG- and GVG-wavegrams.

VI. Acknowledgment

This work is supported by the German Research Foundation (DFG), LO-1413/2.

References

[1] Y. Yan, K. Ahmad, M. Kunduk, and D. Bless, “Analysis of vocal-fold vibrations from high-speed laryngeal images using a hilbert transform-based methodology", *Journal of Voice* 19, 161-175 (2005).
[2] L. Li, N. P. Galatsanos, and D. Bless, “Eigenfolds: a new approach for analysis of vibrating vocal folds", in *ISBI*, 589-592 (2002).
[3] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger, “Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-d diagrams for visualizing and analyzing the underlying laryngeal dynamics", *IEEE Trans. Med. Imaging* 27, 300-309 (2008).
[4] C. T. Herbst, W. T. S. Fitch, and J. G. Švec, “Electroglottographic wavegrams: A technique for visualizing vocal fold dynamics noninvasively", *The Journal of the Acoustical Society of America* 128, 3070-3078 (2010).
[5] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, “Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos", *Medical Image Analysis* 11, 400-413 (2007).
[6] C. Bajaj, V. Pascucci, and D. Schikore, “Data Visualization Techniques”, Volume 6 of Trends in Software, chapter 3: “Accelerated IsoContouring of Scalar Fields”, 31-47 (John Wiley & Sons, Inc.) (1999).
[7] M. Levoy, “Display of surfaces from volume data", *IEEE Comput. Graph. Appl.* 8, 29-37 (1988).
[8] D. Voigt, M. Döllinger, T. Braunschweig, A. Yang, U. Eysholdt, and J. Lohscheller, “Classification of functional voice disorders based on phonovibrograms", *Artificial Intelligence in Medicine* 49, 51-59 (2010).

FULL-AUTOMATIC GLOTTIS SEGMENTATION WITH ACTIVE SHAPE MODELS

J. J. Cerrolaza[1], V. Osma[2], N. Sáenz[2], A. Villanueva[1], J. M. Gutiérrez[2], J. I. Godino[2], R. Cabeza[1]

[1] Department of Electrical and Electronics Engineering, Public University of Navarra, Pamplona, Spain
[2] Department of Circuits and Systems, Polytechnic University of Madrid, Madrid, Spain

Abstract: **In this paper we present a new full-automatic glottis segmentation scheme that combines traditional bottom-up image processing techniques with high-level shape constraints provided by the Active Shape Models. Unlike previous statistical segmentation approaches, which try to accurately detect the location of the glottis as initialization for the algorithm, we incorporate a new reliability score selector at the final stage of the scheme. The result is a robust and flexible algorithm able to deal with most acquisition techniques, even with stroboscopic videos. The good behavior of the algorithm has been successfully tested in a set of 170 frames extracted from 30 stroboscopic recordings.**

Keywords : **Glottis segmentation; Active Shape Models; Region Growing; Stroboscopic videos.**

I. INTRODUCTION

According to the latest studies of the National Institute of Deafness and Communicative Disorders, about 7.5 million individuals suffer from diseases or disorders of the voice due to different causes, such as the overuse of the vocal folds, vocal folds lesions, laryngeal cancer, and other laryngeal pathologies [1]. Although there exists a large variety of techniques for the diagnosis and characterization of these kinds of pathologies, the specialist frequently resorts to visual methods, like the observation of the larynx or the pattern of variation of the vocal folds, to confirm the assessment. In this context, the use of new digital image processing techniques becomes essential to overcome some of the problems inherent to the visualization process, e.g. the presence of artifacts due to non-desired movements of either the patient or the image acquisition system.

One of the aims that has aroused most interest in the research community is the location of the glottal space, whose accurate segmentation is of crucial importance not only to minimize the negative effects of the aforementioned undesirable movements, but also to synthesize and represent the information extracted. Roughly spoken, the different glottal segmentation approaches can be divided into two main groups attending to the underlying philosophy of the algorithm: bottom-up algorithms [2][3], that use image-based criteria to define coherent groups of pixels that belongs to the structure of interest; and top-down methods [4][5], that use information previously learned from a set of examples.

Among the first group, region growing based algorithms are one of the most popular techniques in laryngeal image segmentation. However, in spite of this popularity, their critical initialization dependency makes difficult the full automation of the process, being direct human intervention frequently required. Thanks to the optimal image quality provided by high-speed videos, Wittenberg et al.[2] suggested using the darkest pixel in the image as seed, which is not always correct, especially when working with stroboscopic videos. Also based on high-speed recordings, Chen et al.[3] present an interesting initialization strategy under the assumption that the grayscale histogram of the image follows a mixed Rayleigh distribution, which is not necessary true for stroboscopic images.

One of the first top-down approaches was presented by Saadah et al.[5] using active contours as segmentation strategy, although the initialization remains being an important issue in the proposed framework. As alternative, Demeyer et al.[6] and Skalski et al.[7] use the segmentation obtained in one frame as initialization for the next one. Certainly, this strategy favors the autonomy of the system, although human intervention is still required at the first frame of the video, or if the glottis is obstructed at some point of the recording. One of the first contributions to the glottal segmentation field that incorporated prior shape information using Active Shape Models (ASMs), i.e. statistical shape models built from a set of examples, was the work of Friedl and Wittenberg [4]. The initialization of the algorithm was based on locating the vocal folds motion in a sequence of successive frames extracted from high-speed videos (2000 fps) where the variation between frames is minimal. However, in spite of the goods results and of the high image quality provided by this type of images, the economical cost considerable limits its widespread. Stroboscopic recordings constitute an affordable alternative, although the segmentation problem in this kind of images has aroused little attention.

In this paper we propose a new full-automatic glottal segmentation framework for stroboscopic videos. Unlike other segmentation approaches, this new scheme combines bottom-up techniques, like region growing and basic morphological operations, with high level

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

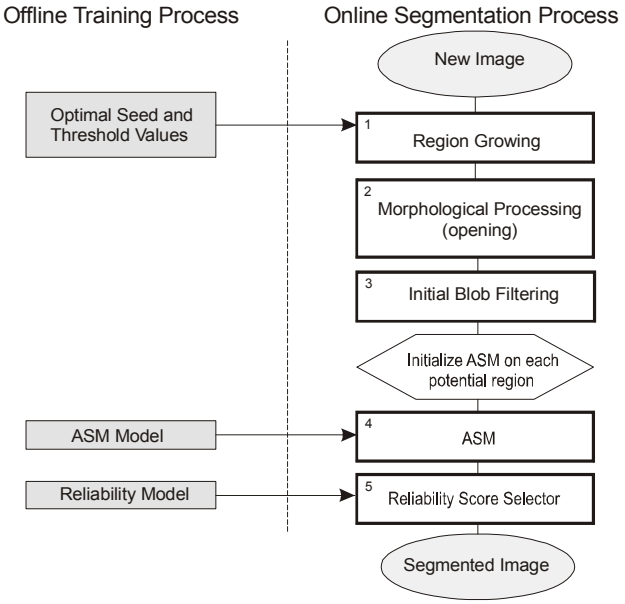


Figure 1. Block decomposition of the new glottis segmentation scheme.

constraints provided by ASMs, creating a new algorithm able to successfully deal with the main drawbacks of stroboscopic videos.

II. METHOD

From a procedural point of view, the new segmentation scheme can be divided into five different blocks or processes, as it is depicted in Fig. 1. When a new image is acquired, an initial coarse segmentation is performed by means of the region growing technique (block 1). The binary image obtained at this initial stage is then processed and filtered by the opening (block 2) and the blog area filtering (block 3) blocks respectively. The target of this initial combination of bottom-up operations is not to obtain an accurate segmentation of the glottis but to provide a set of possible locations of it. The deformable shape model, ASM [8], is initialized at each of these potential positions (block 4), generating one different contour at each of these locations. Making use of the statistical information of appearance and shape used to build the ASM, Sukno and Frangi [9] propose a new reliability score model that will allows us to differentiate the actual glottal area from the rest of fake candidates.

A. Region Growing and Optimal Inputs Parameters

Although a region growing based segmentation approach is not, by itself, accurate enough, it still remains a highly useful initial approximation. However, as in many other pixel-based techniques, the quality of the

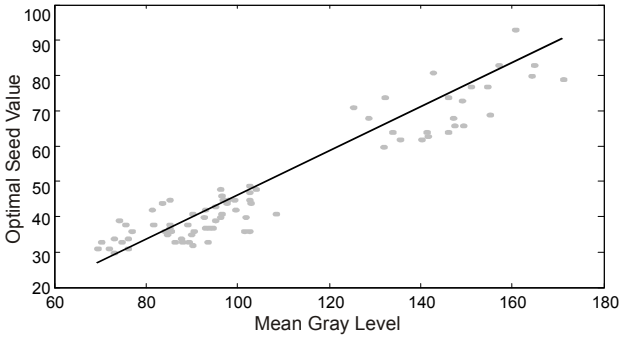


Figure 2. Linear relationship between the average gray level of the image and the optimal seed value of a fixed threshold value of 17 (8-bit grayscale images).

results obtained is strongly dependent on the parameters settings, being the value of the seed the most critical one. The automation of this process turns specially complicated when dealing with stroboscopic images, due to high appearance variability of inter and intra-user images, which makes it particularly hard to define a single default value for the seed or threshold, the two main configuration parameters of the region growing method. As alternative, different combinations of these input parameters, i.e. seed-threshold, are tested over the training set build to create the statistical shape model (ASM). This set is composed of manually segmented examples which allow us to evaluate the suitability of each configuration of input parameters, that is, the region growing segmentation error, and extract the optimal value for each image. Careful experimental tests show how once a default threshold value is fixed, the optimal seed parameter exhibits a strong linear relationship with the average gray level of the input image. Thus, given a new image to segment, it is possible to deduce an adequate seed value by means of the aforementioned relationship with the average gray level of the frame (see Fig. 2).

B. Morphological Processing and Blob Filtering

Once a provisional segmentation has been obtained by means of the process described above (see Fig. 3(b)), it is convenient to include an additional block containing some basic morphological operations, such as opening (i.e. erosion followed by dilation). The erosion operator removes those isolated small regions misclassified during the region growing segmentation, while separates regions mistakenly joined together. On the other hand, the dilation step allows to eliminate undesirable thin protrusions (see Fig. 3(c)) .

Far from obtaining an accurate segmentation of the glottal space, it is worth noting that the goal of this initial processing is to provide a set of potential locations where to initialize the ASM algorithm. Although those invalid locations will be conveniently filtered via the reliability score selector at block 5, it is convenient to reduce the set

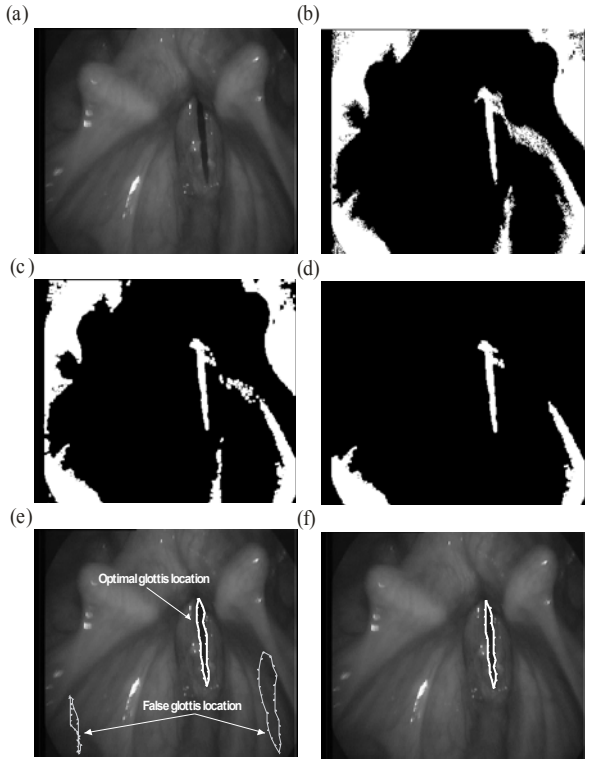


Figure 3. (a) Frame to segment, extracted from a stroboscopic video. (b) Binary image provided by the region growing algorithm. (c) Image after opening. (d) Output of the area filtering. (e) Reliability score result. (f) Final segmentation.

of possible locations as much as possible in order to avoid unnecessary operations. To this purpose, an additional blob filtering block is included. According to our experience a simple area filtering provides satisfactory results for most cases (see Fig. 3(d)), although additional filtering criteria based on the eccentricity or the convexity are also potentially useful.

C. Active Shape Models and their Initialization

Since its presentation in the early nineties by Cootes et al.[8], ASMs have become one of the most popular segmentation paradigms of the last years. A detailed description of the ASM algorithm is out of the scope of this paper, thus the reader is advised to consult the extensive research literature regarding ASM. Roughly speaking, the ASM algorithm can be described as an iterative process in which two statistical models, built from a training set, are sequentially applied to drive the segmentation process. The crucial importance of the training examples is clear, since it must be representative of the variability of both, appearance and shape of the object of interest. In this set of manually segmented images, each shape is defined by a fixed number of points, called landmarks, each one of them defining a specific anatomical point of the shape of interest. The statistical appearance model study the appearance pattern around each landmark, typically the gray pattern of the pixels around it, in order to guides the matching process to a new image. On the other hand, a statistical shape model must to characterize both, the inter-user shape variability, as well as the variability for a particular subject, which corresponds to the vibration of the vocal folds. Once the shape variability has been adequately modeled, global shape constraints can be applied to guarantee that only plausible instances occur.

The aforementioned shape restriction provided by the statistical shape model makes ASMs especially suitable when dealing with anatomic structures, whose variability uses to be conveniently delimited. In the particular case of the glottal area, the statistical model must to characterize both, the inter-user variability, as well as the variability for a particular subject, which corresponds mostly to the patterns of vibration of the vocal folds.

In spite of the satisfactory behavior of ASMs, particularly adequate when modeling anatomical structures, the need for an adequate initialization is still critical. Most of previous solutions [7] concentrate their efforts on identifying an accurate initial location, which typically restrict its applicability to a specific acquisition technique like high speed videos. The alternative proposed in this paper is to use each one of the potential locations obtained at the end of block 3 (see Fig. 1) to initialize the statistical model. The inclusion of a post-filtering additional block will allows us to determine the actual glottal shape.

D. Reliability Score Selector

Combining adequately the information provided by each landmark during the matching process of ASMs, Sukno and Frangi [9] develop a probabilistic framework to obtain a binary evaluation of the reliability of the output. That is, to decide if the shape obtained at the end of the algorithm is reliable (1) or not (0). However, a continuous evaluation of the adequacy of the result is more convenient to our purpose of decide among several potential locations. The global reliability score of a shape, RS, is defined as a weighted average of the individual reliability of all landmarks that defines de contour:

$$RS = \frac{1}{L}\sum_{j=1}^{L}\left(r_j \rho(j \mid j) + \bar{r}_j \rho(j \mid \bar{j})\right) \qquad (1)$$

where L is the number of landmarks; $\rho(j \mid j)$ and $\rho(j \mid \bar{j})$ are the fraction of the j-th landmarks estimated as reliable (r_j = 1) and unreliable ($\bar{r}_j$ = 1) from its appearance model, respectively, for which the landmarks are correctly placed. The reliability of each landmark is evaluated during the matching process, observing its

particular appearance information and deciding if this information is reliable or not according to the statistical model built from the training set. Fixing a threshold of the mahalanobis distance, one of the most popular methods to evaluate the probability that a certain appearance pattern belongs to that observed in the training set, it is possible to decide if a particular location of the landmark is reliable or not. The weighting factors $\rho(j \mid j)$ and $\rho(j \mid \bar{j})$ can be estimated from the segmentation of the shapes contained in the training set, since the actual positions of the landmarks are known.

Computing (1) for each one of those locations where the ASM segmentation algorithm has been initialized, the filtering process is simple: the highest RS value will correspond to the real glottis placement (see Fig. 3(e)).

III. RESULTS

To test the new glottis segmentation scheme presented in this paper, a total of 170 frames have been extracted from 30 stroboscopic videos (25 fps; 360 x 288 pixels) in order to cover the high variability of shapes and illumination conditions observed in the daily practice. The training set is composed by the 70% of these images, using the remaining 30% as testing set.

The behavior of the new full-automatic algorithm is very satisfactory, obtaining an average segmentation error of 5.2 ± 3.4 pixels; while the error obtained when using a manual initialization is of 2.8 ± 2.0 pixels.

It is worth pointing out that the aforementioned test has been performed over isolated frames. However, the new process is amenable of being easily integrated into a frame-by-frame segmentation scheme, using the result at one frame as the initialization for the next one, which will potentially improve the accuracy of the system. In addition, the new reliability score block is a valuable incorporation into the process, allowing us to identify possible inaccuracies (i.e. caused by occlusions) and indicating the need to restart the algorithm from block 1.

IV. CONCLUSION

In this paper we present a new full-automatic segmentation algorithm of the glottal space in stroboscopic images. This work successfully demonstrates how the combination of bottom-up techniques like region growing, with ASMs, one of the latest high level segmentation paradigms, provides an effective solution to the problem, overcoming the two main drawbacks of the existing methods, initialization and acquisition technique dependency.

Using the same statistical information of ASMs, a new reliability score selector has been included. This last block of the segmentation scheme allows us to evaluate each potential location of the target shape and improving the robustness of the system.

The segmentation algorithm has been successfully tested in a set of frames from stroboscopic videos, demonstrating the high potential of the proposed scheme.

REFERENCES

[1] A. E. Aronson, *Clinical voice disorders*. Thieme. 2009.

[2] T. Wittenberg, M. Moser, M. Tigges, U. Eysholdt, "Recording, processing, and analysis of digital high-speed sequences in glottography," *Mach. Vis. Appl.*, vol. 8 (6), pp. 399-404, 1995.

[3] X. Chen, D. Bless, Y. Yan, "A segmentation scheme based on rayleigh distribution model for extracting glottal waveform from high-speed laryngeal images," *27th IEEE-EMBS*, pp. 6269-6272, 2005.

[4] S. Friedl, T. Wittenberg, "Automatic segmentation of vocal folds using active shape models, " *In Procc. Of the 6th Int. Workshop on Adv. In Quantitative Laryngology, Voice and Speech Research*, 2003.

[5] A. K. Saadah, N. P. Galatsanos, D. Bless, C. A. Ramos, "Deformation analysis of the vocal folds from videostroboscopic images sequences of the larynx," *Journal of Acoustical Soc. of America*, 103(6), pp. 3627-3641, 1998.

[6] J. Demeyer, T. Dubuisson, "Glotis segmentation with a high speed glottography: a fully automatic method," *3rd Advanced Voice Function Assessment Workshop*, pp. 113-116, (2009).

[7] A. Skalski, T. Zielinki, D. Deliyski, "Analysis of vocal folds movement in high speed videoendoscopy based on level set segmentation and image registration," *Int. Conf. on Signals and Electronic Systems*, pp. 223-226 (2008).

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, "Active shape models – their training and application," *Comp. Vis. Image Underst*. 64(1), pp. 38-59, (1995).

[9] F. Sukno, A. Frangi, "Reliability estimation for statistical shape models," *IEEE Trans. on Imag. Proc.*, 17(12), pp. 2442-2455, (2008).

ESTIMATION OF GLOTTAL FUNCTIONS USING STEREO-ENDOSCOPIC HIGH-SPEED DIGITAL IMAGING

Ken-Ichi Sakakibara[1], Hiroshi Imagawa[2], Isao T. Tokuda[3],
Hisayuki Yokonishi[2], Miwako Kimura[4], Mamiko Otsuka[5], Niro Tayama[6]

[1] Department of Communication Disorders, Health Sciences University of Hokkaido, Sapporo, Japan
[2] Department of Otolaryngology, The University of Tokyo, Tokyo, Japan
[3] Department of Micro System Technology, Ritsumeikan University, Kusatsu, Japan
[4] Tokyo Voice Center, International University of Health and Welfare, Tokyo, Japan
[5] Kumada Clinic, Tokyo, Japan
[6] National Center for Global Health and Medicine, Tokyo, Japan

***Abstract*: In this paper, a new stereo-endoscopic high-speed digital imaging system and a method to estimate the glottal functions are proposed. Glottal length, width, and area of one male and one female participants were estimated in different fundamental frequencies.**

I. INTRODUCTION

Estimation of glottal functions, such as width, length, and area, plays an important role in clarifying a physical mechanism of vocal fold vibration and investigating voice qualities in a quantitative manner. There have been various methods for estimating glottal area function, however, most of them estimate relative glottal area functions, and actual measurements of glottal area have been done only in vitro.

In this paper, using a method of estimation of time-varying glottal length, width, and area in vivo based on actual measurement by stereoscopic high-speed digital imaging proposed in [1, 2, 3, 4], glottal functions in different F_0s and vocal registers for male and female subjects were estimated. There are positive correlation between F_0 and glottal length, negative correlations between F_0 and glottal width. Shapes of glottal area function varied depending on F_0.

II. METHODS

A. *Stereo-endoscopic high-speed imaging*

The stereo-endoscope includes two independent ordinary rigid optical systems with a diameter of 9 mm. The tips of the optical systems house objective lenses with prisms designed for 70 oblique-angled view, with a field angle of 40° (Fig. 1, 2). The distance between the optical axes of the tips was 10 mm. The stereo-endoscope was attached to a CCTV lens of 50 mm, and the CCTV lens was connected to the high-speed digital camera. The high-speed digital camera employed in this study was Photron Fastcam 1024PCI with the following specifications: an image sensor size of 17.4 mm×17.4 mm, a full image resolution of 1024×1024 pixels, a temporal resolution of 1000 fps at a full image resolution of 1024×1024, 8-bit grayscale.

In stereo-endoscopic high-speed digital recordings, the high-speed camera captured images at an image resolution of 768 (horizontal) ×352 (vertical), a frame rate of 3750 fps, and sample duration of 10.12 s. Fig. 3 shows an example of a pair of stereoscopic images of the larynx. A pair of images was formed side-by-side on the image sensor.

Figure 1: Stereo-endoscope

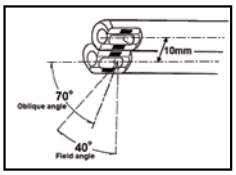


Figure 2: Dimension of stereo-endoscope

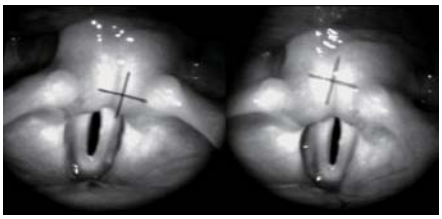


Figure 3: A pair of stereo-endoscopic images of the larynx

B. *Calculation*

Measurements and a procedure of calculation are based on those reported in [1, 2, 3, 4, 7]. The two tips are assumed to be set coplanar, and are mutually inclined to a mid-axis by a small angle α. The distance

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

between the optical axes at the tips is d_T (Figure 4). A rectangular coordinate system is defined with the origin at the tip of the left endoscope. The z-axis is along the optical axis of the left endoscope. The x-axis passes through the two endoscope tips, and the y-axis is orthogonal to the x-axis and the z-axis (out of the page).

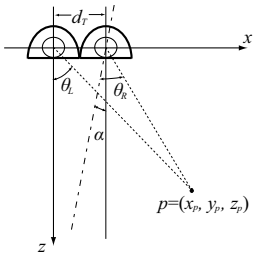


Figure 4: Geometrical quantities defined for calculation of three dimensional coordinates

Vectors to the object point $p = (x_p, y_p, z_p)$ form angles θ_R and θ_L with the left and right optical axes respectively. Let D_L and D_R be horizontal distances of the images of p from the centers of the left and right optical fields respectively, and D_V be a vertical distance of the image of p from the center of the left and right optical fields, then coordinates of p are calculated by the following formulas:

$$z_p = \frac{1}{k_1(D_L - c_1 D_R + c_2 D_V) - k_2} \quad (1)$$

$$y_p = k_3 z_p D_L \quad (2)$$

$$x_p = k_3 z_p D_V \quad (3)$$

where $\{k_i\}$ are calibration constants empirically determined by photographing a Cartesian graph paper. The above calculations are true if D_L and D_R are proportional to $\tan\theta_L$ and $\tan\theta_R$ respectively. In reality, however, a photographic lens causes optical distortion and hence, the relationships such that D_L and D_R are proportional to $\tan\theta_L$ and $\tan\theta_R$ are less likely to be satisfied. Therefore, further calibration and correction of optical distortion are desired. After including correction of optical distortion, the modified formulas to calculate the coordinates are as follows:

$$z_p = \frac{1}{k_1(D_L - c_1 D_R + c_2 D_V) - k_2} \quad (4)$$

$$y_p = k_5 + D_V + k_6 \quad (5)$$

$$x_p = k_3 f(z_p, D_L) + k_4 \quad (6)$$

$$\text{where } f(z_p, D_L) := \frac{D_L - {c_3}^2 + c_4 + z_p + c_5}{c_6 {z_p}^{-c_7}}$$

where $\{k_i\}$ and $\{c_i\}$ are constants for calibration. The procedure to determine constants k_i and c_i was as follows: (i) D_L and D_R were measured by changing distance between the tips of endoscope and the 5 mm Cartesian graph paper from 14 mm to 84 mm; (ii) the regression lines of D_L on x_p and D_R on x_p were calculated for each z_p; (iii) D_L and D_R were represented as functions both having parameters of x_p and z_p; (iv) the regression plane of (D_L, D_R, D_V) was obtained.

As a result, distribution of errors between real coordinates and estimated coordinates in the three-dimensional Euclid space had a median of 0.55 mm ($Q_{0.05} = 0.15$ mm, $Q_{0.95} = 2.96$ mm). The errors of x- and y-axes were less than 15% of the error of z-axis.

C. Glottal edge detection

First, glottal edges in the left and right images both are detected to estimate a glottal area in each frame. On each horizontal line, the edges are automatically determined as the points with maximal brightness derivative among the points with minimum brightness. To represent the glottis as a plane in the three-dimensional space, the following steps were processed: (i) smoothing the estimated left and right edges independently along y-axis by a predetermined window function, reasonably assuming that the edge of glottis is a smooth curve in three-dimensional space, here, the 7-point weighted mean with a length of 7 pixels (0.7 mm at the distance of 50 mm from the endoscope tips in the real space) was employed for smoothing; (ii) determining a regression line of z on y for each edge after the smoothing, and rewriting z in such a manner that the left and right glottal edges were represented as two lines; (iii) for each y, picking up middle point m_y of the left and right glottal edges, then a line approximation C of a curve $\{m_y\}$ was obtained by linear regression of z on y; (iv) calculating lateral inclination from the left edge (x_L, y, z_L) to the right edge (x_R, y, z_R) for each y, and the mean of the lateral inclinations, denoted by $(dz/dx)_{\text{mean}}$; (v) obtaining the glottal hyperplane as the plane including the line approximation C and the hyperplane ' s cotangent vector is orthogonal to $(dz/dx)_{\text{mean}}$. The glottal edge points were obtained as points of projection along z-axis of (x_L, y, z_L) and (x_R, y, z_R) on the glottal hyperplane.

D. Verification of the method

To verify the proposed method for estimation of the glottal edges, the proposed method was applied to estimate a rectangle slit obtained by cut a thick paper (Fig. 5). The rectangle slit had the length of 13.2 mm, the width of 2.2 mm, and the depth of 0.25 mm. Hence, the area of the slit was 29.04 mm^2. Using the proposed method, an area of the glottis was 35.6 mm^2 without smoothing, and 29.5 mm^2 with smoothing and planar approximation. Fig. 6 illustrates the glottis after smoothing in (a) and the estimated glottis after smoothing and planar approximation in (b).

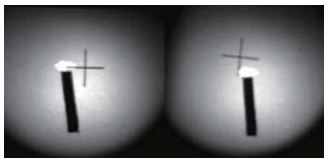


Figure 5: A pair of stereoscopic images of a rectangle slit in a thick paper

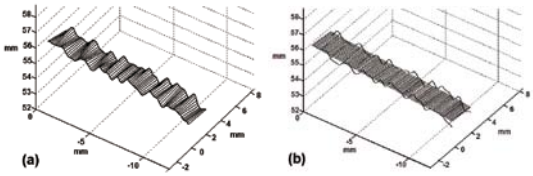


Figure 6: (a): glottis of the slit after edge smoothing and (b): glottis of the slit after smoothing and planar approximation.

III. EXPERIMENTS

One female and one male participants without any vocal problems performed in different F_0s (middle, high, and low), with the same sustained vowel (almost [e] by reason of insertion of endoscope into the mouth). Their vocal fold vibrations were observed by stereo-endoscopic high-speed digital imaging in 3750 fps. A male participant performed in different registers: vocal fry, modal, and falsetto. However, in vocal fry, the vocal folds were covered by supraglottal structures, such as the ventricular and aryepiglottic folds, and not observed by the endoscope, therefore, only modal and falsetto phonations were observed. A female participant performed in three different F_0s and in a modal register.

IV. RESULTS

Fig. 7 shows the glottis after smoothing in (a) and the estimated glottis after smoothing and planar approximation in (b) at $F_0 = 230$ Hz. In this case, the mean lateral inclination $(dz/dx)_{\text{mean}}$ was -0.3. The mean lateral inclinations in the cases of high and low F0s were also equal to -0.3.

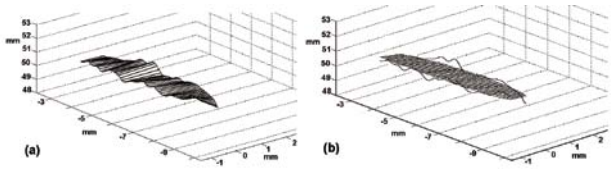


Figure 7: (a): glottis after smoothing and (b): glottis after smoothing and planar approximation in $F_0 = 100$ Hz and a modal register by female.

Figs. 8–10 show a time-varying function of the glottal area (in the top graph), and the glottal width and length (solid and dotted lines, respectively, in the top graph) at $F_0 = 100, 230, 450$ Hz for a female participant. The maximum glottal length increased along with increase of F_0. The maximum glottal length in observed interval was 2.0 mm for 100 Hz, 7.21 mm for 230 Hz, and 8.16 mm for 450 Hz. For the glottal area functions in 100 and 230 Hz, the closing phase was slightly shorter than the opening phase in each period. The maximum glottal widths were 0.85 at 100 Hz, 2.11 at 230 Hz, and 0.92 at 450 Hz.

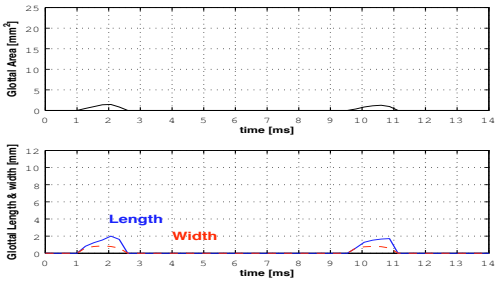


Figure 8: Glottal area (at the top) and glottal length (solid blue line at the bottom), glottal width (dashed red line at the bottom) at $F_0 = 100$ Hz and in modal for female

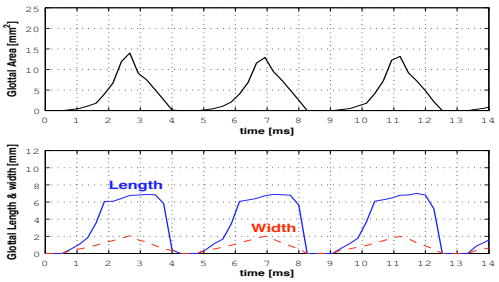


Figure 9: Glottal area, length, and width at $F_0 = 230$ Hz and in modal for female

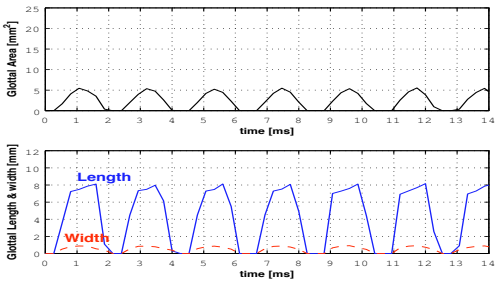


Figure 10: Glottal area, length, and width at $F_0 = 450$ Hz and in modal for female

Figs. 11, 12 show a time-varying function of the glottal area (in the top graph), and the glottal width and length (solid and dotted lines, respectively, in the top graph) at $F_0 = 145$ Hz in a modal register, and $F_0 = 375$ Hz in falsetto register for a male participant.

By observing the glottal area and length functions, the opening phase was slightly longer than the closing phase in a modal register, and the closing phase was longer than the opening phase in a falsetto register. The maximum glottal lengths were 9.14 mm in a modal register at 145 Hz, and 11.07 mm for a falsetto register in 357 Hz. The maximum glottal widths were 2.67 mm in a modal register, and 2.02 mm in a falsetto register.

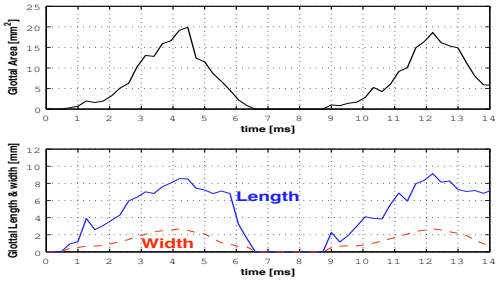


Figure 11: Glottal area, length, and width at $F_0 = 145$ Hz and in modal for male

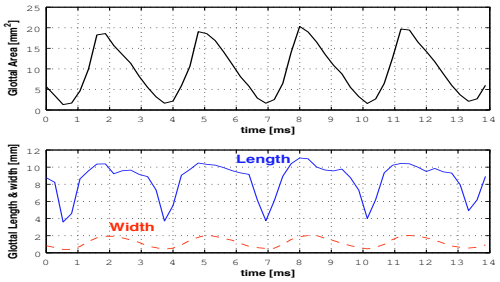


Figure 12: Glottal area, length, and width at $F_0 = 375$ Hz and in falsetto for male

V. DISCUSSION

The glottal functions estimated by the proposed method with stereo-endoscopic high-speed digital imaging were in accordance with known results. The estimated values of the maximum glottal lengths showed good accordance with those in [5, 6]. In the future, it is necessary to improve the method for estimating the glottal area function from the theoretical and instrumental viewpoints.

Acknowledgements: This research was partly supported by Japan and Grant-in-Aid (KAKENHI: 20500161) from the MEXT, Japan and JAIST Grant for Exploratory Research. We would like to thank Kiyoshi Honda for his helpful comments.

REFERENCES

[1] O. Fujimura, T. Baer, and S. Niimi, A stereo-fiberscope with a magnetic interlens bridge for laryngeal observation, *J. Acoust. Soc. Am.*, 65(2):478–480, 1979.

[2] K. Honda, S. R. Hibi, S. Kiritani, S. Niimi, and H. Hirose, Stereoendoscopic measurement of the laryngeal structure, *Ann. Bull. RILP*, 14:73–78, 1980.

[3] H. Imagawa, K.-I. Sakakibara, I. T. Tokuda, M. Otsuka, N. Tayama, Estimation of glottal area function using stereo-endoscopic high-speed digital imaging, *Proc. Interspeech*, 2010.

[4] H. Imagawa, K.-I. Sakakibara, I. T. Tokuda, M. Otsuka, N. Tayama, Estimation of glottal area function using stereo-endoscopic high-speed digital imaging, *J. Phonetic Soc. Jpn.*, 14(2):37–44, 2010, in Japanese.

[5] M. Sawashima, H. Hirose, S. Hibi, H. Yoshioka, N. Kawase, and M. Yamada, Measurements of the vocal fold length by use of stereoendoscope ? a preliminary study, *Ann. Bull. RILP*, 15:9–16, 1981.

[6] M. Sawashima, H. Hirose, K. Honda, H.Yoshioka, S. R. Hibi, N. Kawase, and M. Yamada, Stereoendoscopic Measurement of the Laryngeal Structure, *Vocal Fold Physiology, Contemporary research & Clinical issues* , Edited by Diane M. Bless and James H. Abbs, Collegge-Hill Press, 264–276, 1983.

[7] M. Sawashima and S. Miyazaki, Stereo-fiberscopic measurement of the larynx: a preliminary experiment by use of ordinary laryngeal fiberscopes, *Ann. Bull. RILP*, 8:7–10, 1974. `http://www.umin.ac.jp/memorial/rilp-tokyo/`

Special Session: Computational and experimental vocal fold modelling

Chairperson and introduction:
S. Thomson, C. Brücker

SPECIAL SESSION

COMPUTATIONAL AND EXPERIMENTAL VOCAL FOLD MODELING

Co-chairs: Scott Thomson (*Brigham Young University*),
Christoph Brücker (*TU Bergakademie Freiberg*)

The focus of this session is on recent activities in the areas of computational and synthetic vocal fold modeling. The papers cover a range of topics including material property measurement techniques, reduced-order computational models of esophageal speech, high-fidelity computational models of glottal air flow, and self-oscillating vocal fold models with epithelial layers that yield improved mucosal wave simulation. The submitted abstracts, including authors and their titles and affiliations, are listed below.

Measurement of the elasticity modulus of artificial and real vocal folds using pipette aspiration
S. Weiss (*Dipl.-Ing.*), A. Sutor (*Dr.-Ing.*), J. Ilg (*Dipl.-Ing.*), R. Lerch (*Prof. Dr.-Ing.*)
Chair of Sensor Technology, University of Erlangen, Erlangen, Germany

Parameter optimization for a time-dependent multimass model for the pharyngo-esophageal segment
B. Hüttner (*Dipl.-Phys.*), M. Döllinger (*Prof. Dr.-Ing.*), G. Luegmair (*Dipl.-Ing.*), U. Eysholdt (*Prof. Dr. med. Dr. rer.nat.*), A. Ziethe (*Dr. rer. medic. Dipl. Log.*), and E. Gürlek (*Dr. med.*)
Department of Phoniatrics and Pediatric Audiology, University Hospital Erlangen, Erlangen, Germany

Spectral analysis of the flow in a glottal model
Ch. Brücker (*Prof. Dr.-Ing. habil.*), M. Triep (*Dipl.-Ing.*), C. Kirmse (*Dipl.-Phys.*), W. Mattheus (*Dipl.-Ing.*), R.Schwarze (*Prof. Dr.-Ing.*) Institute of Mechanics and Fluid Dynamics (IMFD), TU Bergakademie Freiberg, Freiberg, Germany

Self-oscillating, multi-layer numerical and artificial vocal fold models with thin epithelial and loose cover layers
S. L. Thomson (*Ph.D.*), P. R. Murray (*B.S.*)
Department of Mechanical Engineering, Brigham Young University, Provo, Utah, USA

MEASUREMENT OF THE ELASTICITY MODULUS OF ARTIFICIAL AND REAL VOCAL FOLDS USING PIPETTE ASPIRATION

S. Weiss[1], A. Sutor[1], J. Ilg[1], R. Lerch[1]

[1] Chair of Sensor Technology, University of Erlangen, Erlangen, Germany

Abstract: **Pipette aspiration technique is applied to take locally resolved measurements of artificial and real vocal folds. The measured data of one-layer and multi-layer samples as well as a pig vocal fold are compared to a finite element simulation in order to estimate the mechanical properties of the investigated samples. For the simulation, we used the results obtained by a previous study. In that study, the mechanical properties of simple one-layer silicone samples have been determined from measurements with a so-called vibration transmission analyzer. With the aid of a mathematical Inverse Method, the numerical results were adjusted to the measured data and thus equations for the material parameters were calculated. These relations serve as the material input parameters for the simulation in our study. Our measurement results are in good agreement with the simulation. Hence, this study verifies the application of an inverse scheme. Moreover, it presents a method for material characterization of multi-layer vocal fold models as well as real tissue that could assist in analyzing voice disorders.**

Keywords: **Elasticity modulus, pipette aspiration, Inverse Method, vocal fold**

I. INTRODUCTION

The quality of life of persons suffering from voice disorders is limited as speech is an important instrument in human communication [1]. It is well known that the vocal folds, more precisely the oscillation of the vocal folds, play an outstanding role in voice generation. Thus, the analysis of vocal fold vibrations could help understanding the reasons of voice disorders. The vibratory characteristics of the vocal folds are mainly influenced by their mechanical material parameters, especially the elasticity modulus [2]. This fact demands for methods to estimate the mechanical properties to describe the vocal fold vibration and thus assist in improving the clinical care of human voice.

Many different approaches to characterize the mechanical material parameters of the vocal folds have been published. In [3,4] the static elasticity modulus was determined by analyzing the stress relaxation of canine vocal folds with the aid of an ergometer. To measure the dynamic mechanical properties of viscoelastic materials, rheometer systems were applied in [5,6]. With these systems, the dynamic material behavior up to 250 Hz could be determined. Furthermore, they are destructive methods and very expensive. A low-cost alternative to identify the mechanical material parameters within the frequency range of human phonation was presented in [7]. In that study, the transfer function of cylindrically shaped one-layer silicone samples was measured with a so-called vibration transmission analyzer. For the estimation of the mechanical properties an Inverse Method [8] is applied, that minimizes the deviations between a simulated transfer function and the measured one by adjusting the sought-after parameter set. The comparison of numerical and measured results showed only small deviations. However, this method is only applicable for homogeneous one-layer samples. Thus, the idea of this study is to find a possibility to determine the dynamic material parameters of multi-layer samples on the basis of the results presented in [7].

In this study, the pipette aspiration technique [9,10] is applied to measure the local stiffness of artificial and real vocal folds. With this technique, the elasticity modulus is determined by the measurement of the maximum aspiration displacement, under assumption that the investigated specimen is isotropic, incompressible, homogeneous and linearly elastic. Realistic vocal fold models are usually made of silicone mixtures [11,12] that fulfill the required properties. Therefore, we investigated several one-layer and multi-layer silicone samples with reference to their maximum aspiration displacement as well as to their displacement profiles within the human voice frequency range. The measurement data is compared to a finite element simulation of the experimental setup which is based on the adjusted dynamic material parameters out of the investigations with the vibration transmission analyzer [7]. In order to compare the artificial specimens with real tissue, a pig vocal fold has also been investigated.

II. METHODS

A. Measurement Setup

The idea of determining the elasticity modulus using pipette aspiration was first reported in [9]. To analyze the dynamic material behavior, the original setup was extended [13] and has provided the basis for the resulting pipette aspiration setup in this study, which is shown in

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Fig. 1. The pipette is centrically placed on the surface of the specimen. To create boundary conditions that are in good agreement with the simulation, the pipette is glued on the silicone samples. Concerning the investigations of the pig vocal fold, the contact force is measured by a force sensor and kept constant at 0.1 N. The pistonphone, which is connected to the gauge head by a flexible tube, is mounted on an electromechanical shaker. This shaker induces a mechanical oscillation, the frequency and the amplitude of which are controlled by a computer via the automation software LabVIEW©. This oscillation is transmitted to the lobe of the pistonphone. Thus, a fluctuating pressure is generated in the gauge head resulting in a vibration of the area enclosed by the pipette. An electret microphone (SENNHEISER© KE4-211-2) detects the actual pressure in the gauge head and by means of a controller the pressure can be maintained. The laser scanning vibrometer (PSV 300 from POLYTEC©), communicating with the control computer, scans defined points on the specimen's surface and measures the out-of-plane velocity at each point. In order to avoid measurement errors due to outside vibrations the setup is placed upon a vibration-isolated table.

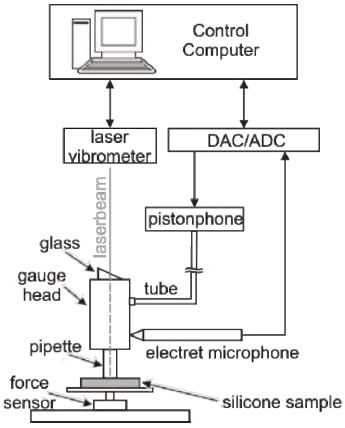


Fig. 1: Pipette aspiration setup

B. Specimens

The conventional material for modeling vocal folds is silicone rubber as its material properties are comparable to those of real vocal folds [11,12]. The samples in our investigations were made of the three-component addition-cure silicone Ecoflex 0030© (Smooth-On, Inc.), consisting of equal amounts of subcomponents part A and part B as well as a variable amount of silicone thinner T. The sample's stiffness depends on the used amount of silicone thinner. The smaller the amount of silicone thinner, the higher is the elasticity modulus. Three different mixtures (A:B:T) are investigated in this work: same amounts of subcomponents and thinner (1:1:1), twice as much thinner compared to the subcomponents (1:1:2) and mixtures with three parts of thinner (1:1:3). With regard to a simple simulation model of the silicone samples, our specimens are cylindrically shaped with d = 50 mm in diameter and h = 10 mm in height.

As real vocal folds consist of several layers with different mechanical material parameters [14], multi-layer samples have also been fabricated and analyzed. A two-layer sample has been taken into account. The base is a layer of a 1:1:1-mixture with 5 mm in height. This layer is covered with a 2 mm thick layer of a 1:1:2-mixture. Furthermore, a three-layer sample with an additional third layer of a 1:1:3-mixture with 1 mm in height has also been investigated.

In order to compare the frequency dependent behavior of the silicone samples with real tissue, one vocal fold of a pig larynx has been excised and examined.

III. RESULTS

A. Investigations

All investigations in this study were performed by using a pipette with an inner radius of r = 3 mm. In order to cover the frequency range of human phonation, an interval from 50 to 300 Hz is considered. The applied pressure for the specimens is adjusted by a controller to 20 Pa over the whole frequency range. With the laser scanning vibrometer, the amplitudes of the velocity of about 70 points on the surface of the investigated samples have been measured at frequency intervals of $\Delta f = 1$ Hz. The displacement is determined by integrating the velocity over time.

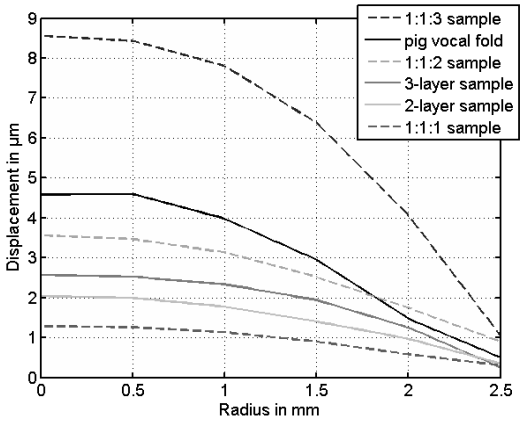


Fig. 2: Comparison of the displacement profiles for the different specimens at 120 Hz

Figure 2 exemplary displays the mean of the measured displacements at discrete positions within the pipette at a frequency of 120 Hz. Due to reflections of the pipette wall, the radial position is only plot until 2.5 mm. The displacement profile of the stiffest sample (1:1:1 sample) shows a flat decay towards the inner pipette wall whereas the 1:1:3 sample shows the steepest slope. The

maximal displacement of the investigated multi-layer samples lies between that of the 1:1:1 sample and that of the 1:1:2 sample. With an inner pipette radius of 3 mm, this is obvious as the layer of the stiffest material is the thickest of the specimen and thus mainly determines the dynamic behavior of the sample. The maximum displacement of the pig vocal fold at 120 Hz is ca. 4.5 µm, which is in between the maximum of the samples with mixing ratios of 1:1:2 and 1:1:3. This indicates that the real tissue has a different layer structure than the investigated multi-layer silicone samples.

B. Finite Element Simulation

In order to compare the measurement results to the results of a finite element simulation, a finite element model (FEM) has been designed using ANSYS© preprocessor. The simplified axially symmetric 2D-model of the three-layer cylindrical specimen is shown in Fig. 3. The bottom nodes' motion as well as the region on which the pipette is placed on the sample are fixed in all directions. A constant pressure of 20 Pa is applied to the aspiration area. The different layers consist of materials that are assumed to be homogeneous and isotropic but with different elasticity modulus, damping factor and Poisson's ratio. The input values for the dynamic material parameters of the different silicone mixtures are extracted from [7]. The mechanical displacements at 16 equal distant nodes on the aspiration area are calculated. The numerical simulations are performed with the finite element software package CFS++ (Coupled Field Simulation) [15] within a frequency range of 50-300 Hz with 100 linearly distributed steps.

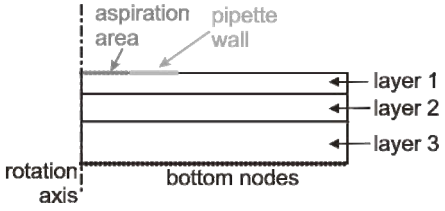


Fig. 3: Simplified rotationally symmetric 2D-FEM of the three-layer sample

C. Comparison of measurement and simulation

The results of the maximum displacements as a function of frequency for the investigated specimens are displayed in Fig. 4. The continuous curves show the measurement results, the dashed lines show the simulated characteristics. In general, the more elastic the material is the lower is the resonance frequency. For all silicone samples, the resonance frequency of both the measurement and the simulation are approximately identical. Furthermore, the absolute values of the measured curves agree with the simulated ones.

However, close to the resonance frequency, the measured displacements are about 1-2 µm higher than the simulated ones. One possible reason for this may be that the region on which the pipette is placed on the sample is not totally fixed in all directions as a loss in fixation has been detected during the measurement. Moreover, material parameters calculated in [7], which serve as input for the simulation, have not been determined for the samples investigated in this study. Thus, a minor error due to the differences in the set simulation values for the damping factor and the complex elasticity modulus occurs. The measurement results of the multi-layer samples also show good agreement with the simulation. Regarding the results of the pig vocal fold, the resonance frequency is much lower than that of the multi-layer samples.

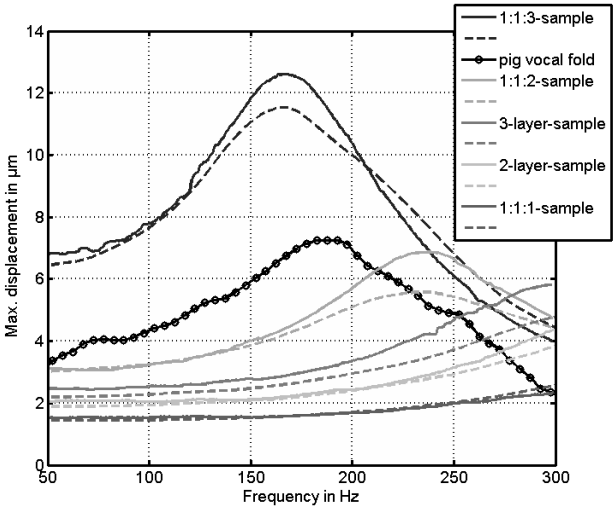


Fig. 4: Measured and simulated maximum displacements over frequency for the different samples

IV. Discussion

The presented results show that the pipette aspiration method can be used for locally resolved measurements of the mechanical properties of vocal folds. The maximum displacements and the displacement profiles of one-layer and multi-layer silicone samples as well as an excised pig vocal fold have been measured. It has been shown that for the measurements with a pipette with an inner radius of 3 mm, the maximum displacements and thus the dynamic behavior of multi-layer samples are mainly determined by the thickest layer. In order to analyze the influence of the cover layers, further investigations will be made using a pipette with a smaller inner radius. In comparison to the measured characteristics of real tissue, the investigated silicone specimens show a much higher resonance frequency. To approach the frequency response of real vocal folds, the chosen layer thicknesses have to be

adjusted. But in general, the results prove the possibility to model a silicone sample whose behavior is close to that of real tissue. However, for this study the dynamic behavior of only one pig vocal fold has been measured. For further investigations, several pig vocal folds have to be examined so that the reproducibility of the results can be guaranteed.

A comparison with a finite element simulation on the basis of the mechanical properties calculated in [7] shows that the absolute values as well as the resonance frequencies of measured and simulated data are approximately identical. This fact verifies the determination of the dynamic material parameters with the Inverse Method. Consequently, future researches will concentrate on an application of the Inverse Method to the presented pipette aspiration setup. Thus, our method could be used to determine the complex elasticity modulus, the damping factor and the Poisson's ratio of investigated multi-layer samples and help in modeling the dynamic behavior of real tissue.

V. Conclusion

A simulation based method to determine the elasticity modulus of artificial and real vocal folds has been presented. Pipette aspiration technique is applied to take locally resolved measurements of silicone samples and a pig vocal fold. The measurement data is compared to a finite element simulation. The results of the estimation of dynamic mechanical properties by an Inverse Method [7] serve as input material parameters. The comparison of measurement and simulation shows accordance concerning the frequency responses of the investigated silicone samples. The examination of the pig vocal fold showed a much lower resonance frequency than that of the artificial multi-layer samples, indicating that the thickness ratios have to be adjusted to approach the behavior of real tissue.

The advantage of our method to tensile tests [3,4] or rheometer systems [5,6] is that it is nondestructive. In comparison to the material characterization of one-layer samples with a vibration transmission analyzer [7], the benefit of the presented method is the possibility to determine the mechanical properties of multi-layer samples, the structure of which is similar to that of real vocal folds. Moreover, this method offers locally resolved measurements and could thus be a possible mean in improving the clinical care of human voice.

References

[1] M. Doellinger, "The next step in voice assessment: High-speed digital endoscopy and objective evaluation", *Current Bioinformatics*, vol. 4, pp. 101-111, 2009

[2] I. R. Titze and D. T. Talkin, "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation", *Journal of the Acoustical Society of America*, vol. 66, pp. 60-74, 1979

[3] F. Alipour-Haghighi and I. R. Titze, "Viscoelastic modeling of canine vocalis muscle in relaxation", *Journal of the Acoustical Society of America*, vol. 78, pp. 1939-1943, 1985

[4] F. Alipour-Haghighi and I. R. Titze, "Elastic models of vocal fold tissues", *Journal of the Acoustical Society of America*, vol. 90, pp. 1326-1331, 1991

[5] I. R. Titze, S. A. Klemuk and S. Gray, "Methodology for rheological testing of engineered biomaterials at low audio frequencies", *Journal of the Acoustical Society of America*, vol. 115, pp. 392-401, 2004

[6] R. W. Chan and M. L. Rodriguez, "A simple-shear rheometer for linear viscoelastic characterization of vocal fold tissues at phonatory frequencies", *Journal of the Acoustical Society of America*, vol. 124, pp. 1207-1219, 2008

[7] S. J. Rupitsch, J. Ilg, A. Sutor, R. Lerch and M. Doellinger, "Simulation based estimation of dynamic mechanical properties for viscoelastic materials used for vocal fold models", *Journal of Sound and Vibration*, vol. 330, pp. 4447-4459, 2011

[8] S. J. Rupitsch and R. Lerch, "Inverse method to estimate material parameters for piezoceramic disc actuators", *Applied Physics A: Materials Science & Processings*, vol. 97, pp. 735-740, 2009

[9] T. Aoki, T. Ohashi, T. Matsumoto and M. Sato, "The Pipette Aspiration Applied to the Local Stiffness Measurement of Soft Tissues", *Annals of Biomedical Engineering,* vol. 25, pp. 581-587, 1997

[10] T. Ohashi, H. Abe, T. Matsumoto and M. Sato, "Pipette aspiration technique for the measurement of non-linear and anisotropic mechanical properties of blood vessel walls under biaxial stress", Journal of Biomechanics, vol. 38, pp. 2248-2256, 2005

[11] S. L. Thomson, L. Mongeau and S. H. Frankel, "Aerodynamic transfer of energy to the vocal folds", *Journal of the Acoustical Society of America*, vol. 118, is. 3, pp. 1689-1700, 2005

[12] D. A. Berry, Z. Zhang and J. Neubauer, "Mechanisms of irregular vibration in a physical model of the vocal folds", *Journal of the Acoustical Society of America,* vol. 120, is. 3, pp. 36-42, 2006

[13] S. Zoerner, M. Kaltenbacher, A. Sutor and M. Doellinger, „Measurement of the elasticity modulus of soft tissues", *Journal of Biomechanics*, vol. 43, pp. 1540-1545, 2010

[14] M. Hirano, "Structure and vibratory behavior of the vocal fold", in *Dynamic Aspects of Speech Production,* edited by M. Sawashima and F .S. Cooper (University of Tokyo, Tokyo), pp. 13-30, 1977

[15] M. Kaltenbacher, *Numerical Simulation of Mechatronic Sensors and Actuators*, 2nd Ed., Springer, Berlin, 2007

PARAMETER OPTIMIZATION FOR A TIME-DEPENDENT MULTI- MASS MODEL FOR THE PHARYNGO-ESOPHAGEAL SEGMENT

B. Hüttner[1], M. Döllinger[1], G. Luegmair[1], U. Eysholdt[1], A. Ziethe[1], and E. Gürlek[1]

[1] Department of Phoniatrics & Pediatric Audiology, University Hospital Erlangen, Erlangen, Germany

Abstract: **Laryngeal cancer may necessitate a complete removal of the larynx. Consequently, the required sound source for voiced communication is lost. Alternatively, a substitute sound signal can be generated by tissue vibrations in the pharyngo-esophageal (PE) segment. The quality of the substitute voice significantly depends on the vibration characteristics of the PE segment. For investigation purpose, the tissue vibrations are detected by endoscopic interventions with a high-speed camera and are quantified by a biomechanical multi-mass model. As pseudoglottal vibrations present variations in frequency and amplitude we suggest an expansion of the biomechanical PE model for a time-dependent multi-mass model. Additionally, we propose a block based optimization procedure to fit the model dynamics to real PE vibrations. First results demonstrate the performance of the time-dependent model and the optimization procedure being comparable to those of the non-stationary PE model and the time-dependent vocal fold model.**

Keywords: **PE segment, model optimization**

I. INTRODUCTION

The tasks of the larynx are the separation of the esophagus and the trachea as well as the generation of the sound source necessary for voiced communication. Due to cancer a *laryngectomy* may be necessary, i.e. removal of the complete larynx at which laryngeal functions are lost. However, the functions can be "reconstructed" by a surgical intervention (Fig. 1) [1]. At first the esophagus and the trachea are separated. To preserve breathing, the trachea is sewn into a respiratory notch in the frontal neck, the so called *tracheostoma*. The pharynx is directly connected with the esophagus at which the scared tissue in the changeover is called *PE segment* (pharyngo-esophageal segment). To allow for a substitute sound source, a valve is inserted to connect trachea and esophagus. Closing the tracheostoma forces the air stream from the lungs to pass the valve and the PE segment. The scarred tissue is stimulated for oscillation what generates the sound source for *tracheo-esophageal voice production* [2].

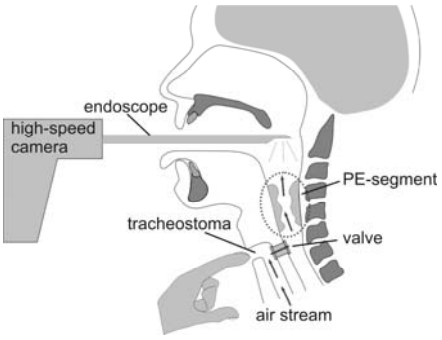


Fig 1: Endoscopic examination setup of the PE vibrations by a high-speed camera. Closing the tracheostoma forces the air stream from the lungs to pass the valve and stimulate the tissue in the PE segment for oscillation.

The intelligibility of the tracheo-esophageal substitute voice is drastically reduced compared to a healthy voice [3]. Moreover a broad variability in quality exists [4]. The latter is mainly determined by the vibration patterns of the PE-segment [5]. For a quantitative analysis, the oscillations of the PE-segment are recorded by a high-speed camera (Fig. 1, Fig. 2). The time-signal of the opening area of the PE-segment (*pseudoglottis*) is extracted (Fig. 2) and is modeled by a stationary biomechanical multi-mass model (PE-MMM) [6]. However, the time-signal presents variations in amplitude and frequency, see Fig. 2. Thus we suggest the expansion of the PE-MMM to a time dependent model (PE-MMM(t)) by applying time dependent model parameters. For the simulation of real PE vibrations, the model parameters of the PE-MMM(t) are fitted by a block based optimiza-

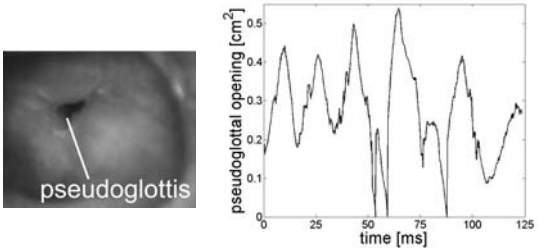


Fig 2: Left: Endoscopic view of the PE segment. The black bead is the pseudoglottis. Right: Time signal of the pseudoglottal opening area.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

tion procedure to the non stationary time signal of the pseudoglottal opening extracted of the high-speed recordings. The resulting model parameters objectively quantify the vibrations of the PE-segment.

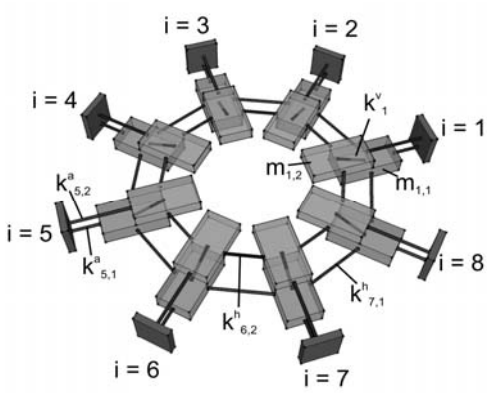


Fig 3: PE-MMM(t) consisting of i=8 vertically and horizontally coupled mass-spring oscillators $m_{i,s}$. $k^a_{i,s}$, $k^v_{i,s}$ and $k^h_{i,s}$ are the anchor, the vertical and the horizontal couplings, respectively, of the lower (s=1) and upper (s=2) plane. Damping elements are omitted.

II. METHODS

Alike to the stationary PE-MMM, the pseudoglottal contour is modeled by 8 two-mass oscillators [6] that are coupled by horizontal and vertical springs k^h, k^v and dampers r^h, r^v. The masses are arranged in a horizontal plane and a closed shape (Fig. 3). The masses are additionally connected by horizontal couplings (k^a, r^a) to anchors with fix positions. Each mass is capable to move within the whole plane at which its position at a discrete time step n is described by the 2D vector $\mathbf{x}_{i,s}(n) = (x_{i,s}(n), y_{i,s}(n))^T$. The index $i = 1,\ldots,8$ describes the i^{th} mass element in the lower (s=1) and upper (s=2) plane, respectively. The time-dependent parameters of the mass-spring oscillations are the masses $m_{i,s}(n)$, their rest positions $\mathbf{x}^r_{i,s}(n)$, the horizontal anchor couplings $k^h_{i,s}(n)$, and the subglottal pressure $P^{sub}(n)$. The trajectories $\mathbf{x}_{i,s}(n)$ of the mass elements $m_{i,s}$ obey the 2nd Newtonian Law:

$$m_{i,s}\ddot{\mathbf{x}}_{i,s} = -\dot{m}_{i,s}\dot{\mathbf{x}}_{i,s} + \mathbf{F}^a_{i,s} + \mathbf{F}^v_{i,s} + \mathbf{F}^h_{i,s} + \mathbf{F}^c_{i,s} + \mathbf{F}^d_{i,s}. \quad (1)$$

$\mathbf{F}^a_{i,s}$, $\mathbf{F}^v_{i,s}$ and $\mathbf{F}^h_{i,s}$ are the anchor, the vertical and the horizontal coupling force. $\mathbf{F}^c_{i,s}$ is the force due to collisions with other masses or horizontal coupling springs. The force $\mathbf{F}^d_{i,s}$ is the driving force generated by the glottal airflow caused by the subglottal pressure $P^{sub}(n)$ [6].

For the simulation of real PE vibrations the contour and the area signal of the pseudoglottis are extracted from high speed recordings (HSR) [4]. Afterwards the area and contour of the model are fit to the extracted data. Thereto, the model parameters are adjusted by sets of 10 time dependent optimization parameters $P(n) = [Q_1(n),\ldots,Q_8(n),Q_P(n),Q_r(n)]$ [6]:

$$\begin{aligned} m_{i,s}(n) &= \hat{m}_{i,s} / Q_i(n), \\ k^a_{i,s}(n) &= \hat{k}^a_{i,s} \cdot Q_i(n), \\ P^{sub}(n) &= \hat{P}^{sub} \cdot Q_P(n), \\ \mathbf{x}^r_{i,s}(n) &= \mathbf{p}(n) + (\hat{\mathbf{x}}^r_{i,s} - \mathbf{p}(n)) \cdot Q_r(n). \end{aligned} \quad (2)$$

The 'hats' indicate constant initialization values. A cost function

$$\Gamma = (\Delta a)^2 + (\Delta s)^2 + (d)^2 \quad (3)$$

consisting of three minimization criteria is defined to find adequate Q-values [6].

1) **Area consistency:**

$$\Delta a = \frac{\sum_{n=1}^{N} \left| a^{\text{mod}el}(n) - a^{HS}(n) \right|^2}{\sum_{n=1}^{N} \left| a^{HS}(n) \right|^2}, \quad (4)$$

with $a^{model}(n)$ and $a^{HS}(n)$ being the glottal area generated by the model and extracted from HSR at the discrete time step n, respectively.

2) **Intersection consistency:**

$$\Delta s = \frac{\sum_{n=1}^{N} (|\, a^{\text{mod}el}(n) - a^{\text{sec}}(n)\,| + |\, a^{HS}(n) - a^{\text{sec}}(n)\,|)}{\sum_{n=1}^{N} (|\, a^{\text{mod}el}(n) + a^{HS}(n)\,|)}, \quad (5)$$

with $a^{sec}(n)$ being the intersection area between $a^{model}(n)$ and $a^{HS}(n)$.

3) **Time-averaged distance:**

For each time step the minimal distance $d_i(n)$ between mass $m_{i,s}$ and the extracted contour is determined. $d_i(n)$ is normalized to the radius of a circle with same area as $a^{HS}(n)$:

$$r(n) = \sqrt{a^{HS}(n)/\pi}\,. \quad (6)$$

$$d = \frac{1}{8N} \sum_{i=1}^{8} \sum_{n=1}^{N} \frac{d_i(n)}{r(n)}. \quad (7)$$

A block based optimization procedure is applied to minimize the objective function Γ [7]. The area signals $a^{HS}(n)$ and $a^{model}(n)$ are split into blocks, each containing four oscillation cycles at which a cycle is defined from maximum to maximum. To assure smoothness, consecutive blocks have an overlap of 50%. Γ is minimized in each block by a combination of *Adaptive Simulated Annealing* and *Powell's direction set method*. Though the complete block is minimized, the optimized parameters of the second half of the block are rejected as they are optimized in the first half of the consecutive block. The resulting parameter sets P(n) are applied as model

Tab 1: Relative errors of Q_i, Q_p and Q_r, averaged over the five runs, and the predefined Q-values of P^*_1 and P^*_2, respectively.

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_P	Q_r
mean relative error P^*_1 [%]	6.7	4.5	9.3	9.0	9.4	7.9	10.4	15.5	9.6	8.7
mean relative error P^*_2 [%]	6.1	5.0	7.0	6.5	4.4	7.6	7.2	7.6	4.6	6.9

parameters for the PE-MMM(t) to generate the time-dependent pseudoglottal openings at every time step *n*.

The reliability and the capability of the optimization procedure are validated by applying the suggested method to synthetically generated data. Thereby, the generated model dynamics of predefined parameter sets $P^*(n)$ serve as presetting to be estimated by the optimization procedure. Two kinds of parameter sets are applied. P^*_1 varies Q_p and P^*_2 varies Q_i by linearly increasing the concerning Q-value over time. All other Q values of P^*_1 and P^*_2 remain constant. The optimization procedure was applied five times to each parameter set.

To demonstrate the applicability, the optimization procedure is applied to two PE vibrations extracted from HSR.

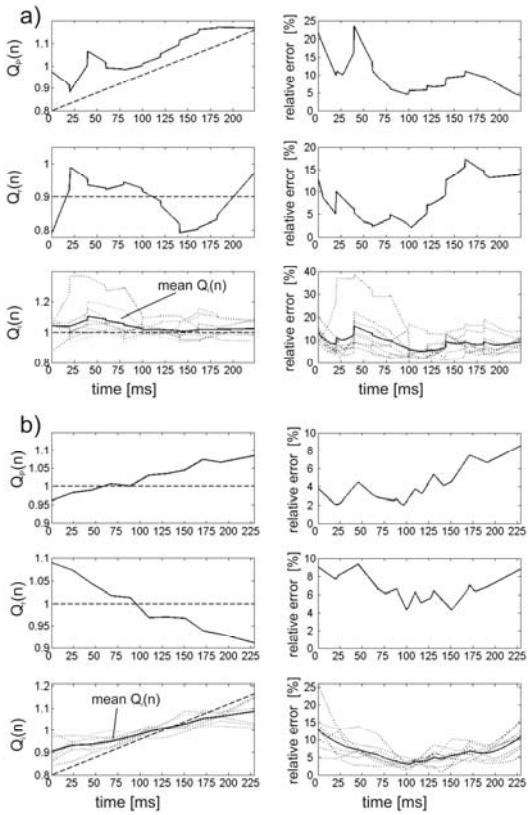


Fig 4: Optimization results of fitting the PE-MMM(t) to the data set of P^*_1 (a) and P^*_2 (b). Left column: Q-values over time steps. The dashed lines are the predefined Q-values, the solid lines are the average of the optimized Q-values over the five runs. The right column depicts the relative error between the predefined values and the averaged optimized values.

III. Results

A. Validation by synthetic data sets

The optimization results after fitting the PE-MMM(t) to the time signal of the synthetic data sets P^*_1 and P^*_2 are depicted in Fig. 4a) and 4b). The left columns show the course of Q_P, Q_r and Q_i ($i = 1 \ldots 8$) over time. The dashed lines are the predefined Q-values of P^*, the solid lines are the optimized Q-values averaged over the five runs. The right columns present the relative error.

The means of the relative errors are summarized in Tab. I. The mean errors of Q_i have values between 4.5% and 15.5% with an average of 9.1% for P^*_1, and values between 4.4% and 7.6% with an average of 6.4% for P^*_2. The average over all Q-values is 9.1% for P^*_1 and 6.0% for P^*_2.

B. Application to real PE vibrations

Fig. 5 and Fig. 6 show the optimization results after fitting the PE-MMM(t) to real PE vibrations extracted from HSR. Fig. 5a) and 6a) depict the time signals of the pseudoglottal area. During the optimization the signal of the PE opening is split into blocks with four oscillation cycles, at hand three blocks. The optimization results of blocks 1 to 3 are depicted in Fig. 5b)-d) and Fig. 6b)-d). The solid line is the time signal of the PE area, the dashed line the area of the optimized PE-MMM(t). The correlations for both curves in blocks 1 to 3 are 98.2%, 97.8%, 98.5% for Fig. 5 and 98.0%, 99.7%, 99.7% for Fig. 6.

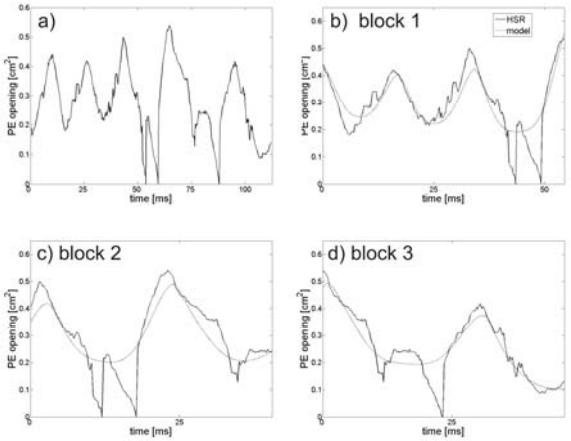


Fig 5: Optimization result of fitting the PE-MMM(t) to real PE oscillations extracted from HSR. a) Time signal of the PE opening area. b) to d) Optimization results within the individual blocks of the optimization procedure. The solid line is the time signal of the pseudoglottal area, the dashed line is the time signal of the optimized model opening.

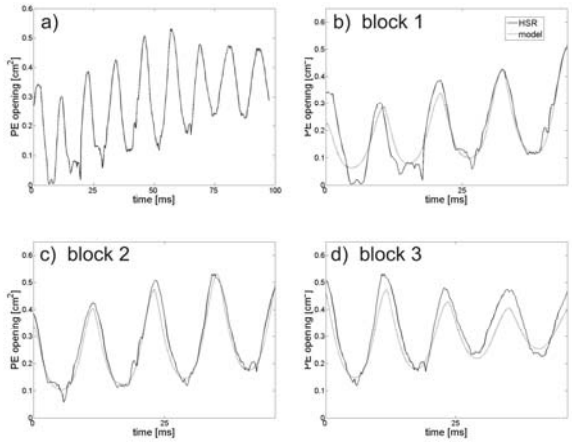


Fig 6: Optimization result of fitting the PE-MMM(t) to real PE oscillations extracted from HSR. a) Time signal of the PE opening area. b) to d) Optimization results within the individual blocks of the optimization procedure. The solid line is the time signal of the pseudoglottal area, the dashed line is the time signal of the optimized model opening.

IV. Discussion

To demonstrate the capability of the suggested optimization procedure the model dynamics of the two predefined parameter sets P^*_1 and P^*_2 were both optimized five times. The optimized Q-values present relative errors in the range of 4.5% and 15.5% for P^*_1 and 4.4% and 7.6% for P^*_2 with averages over all ten Q-values of 9.1% and 6.0%. The results show that the optimization works better in optimizing variations in Q_i than variations in Q_P. Optimizations applying the stationary PE-MMM show relative errors in the range of 4.9% to 12.3% [6]. Our results are in the same order of magnitude. The mean error for the PE-MMM amounts for 8.2% and lies between the here presented mean errors for P^*_1 and P^*_2. In comparison to the non-stationary multi mass model for the vocal folds that results in a mean error for the Q-values of 10.9% [7], the here presented PE-MMM(t) achieves better optimization results.

To demonstrate the applicability of the PE-MMM(t) and the suggested optimization procedure, two area functions of real PE vibrations were extracted from HSR. The optimization resulted in correlations of the area functions of the model and the PE segment in the range of 97.8% - 99.7%. Correlations while applying the stationary PE-MMM amounts to 69% - 95% [6]. The non-stationary model of the vocal folds achieves correlations between 89% and 97% [7]. The results demonstrate the capability of the suggested PE-MMM(t) to represent vibrations of the PE segment.

The future application of the PE-MMM(t) will be the objective quantification of PE vibrations what is represented by the model parameters after fitting the model dynamics to the PE oscillations. These findings may help to improve voice rehabilitation by detecting regions that have significant influence to the quality of the substitute voice. This knowledge can be used to improve the surgical intervention during laryngectomy to optimize the scarring of the PE segment.

V. Conclusion

First optimizations show that the performance of the suggested PE-MMM(t) is in good agreement with that of the stationary PE-MMM [6] and that of the non-stationary multi-mass model for the vocal folds [7].

Acknowledgement

The work was supported by 'Deutsche Krebshilfe' grant no. 109204: "Analyse und Modellierung der pharyngoösophagealen Schleimhautdynamik nach krankheitsbedingter Kehlkopfentfernung".

References

[1] J.D. Dahm, D.G. Sessions, R.C. Paniello and J. Harvey, "Primary subglottic cancer", *Laryngoscope*, vol. 108, 1998.

[2] B. Elmiyeh, R.C. Dwivedi, N. Jallali, E.J. Chisholm, R. Kazi, P.M. Clarke and P. H. Rhys-Evans, "Surgical voice restoration after total laryngectomy: an overview", *Indian J. Cancer*, vol. 47, 2010.

[3] M. Schuster, J. Lohscheller, P. Kummer, U. Hoppe, U. Eysholdt and F. Rosanowski, "Quality of life in laryngectomees after prosthetic voice restoration", *Folia Phoniatr. Logop.*, vol. 55, 2003.

[4] J. Lohscheller, M. Döllinger, M. Schuster, R. Schwarz, U. Eysholdt and U. Hoppe, "Quantitative investigation of the vibration pattern of the substitute voice generator", *IEEE Trans. Biomed. Eng.*, vol. 51, 2004.

[5] M. Schuster, R. Rosanowski, R. Schwarz, U. Eysholdt and J. Lohscheller, "Quantitative detection of substitute voice generator during phonation in patients undergoing laryngectomy", *Arch. Otolaryngol. Head Neck Surg.*, vol. 131, 2005.

[6] R. Schwarz, B. Hüttner, M. Döllinger, G. Luegmair, U. Eysholdt, M. Schuster, J. Lohscheller and E. Gürlek, "Substitute Voice Production: Quantification of PE Segment Vibrations Using a Biomechanical Model", *IEEE Trans. Biomed. Eng.*, accepted 2011.

[7] T. Wurzbacher, M. Döllinger, R. Schwarz, U. Hoppe, U. Eysholdt and J. Lohscheller, "Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters", *J. Acoust. Soc. Am.*, vol. 123, 2008.

SPECTRAL ANALYSIS OF THE FLOW IN A GLOTTAL MODEL

Ch. Brücker , M. Triep , C. Kirmse , W. Mattheus , R. Schwarze

Institute of Mechanics and Fluid Dynamics (IMFD), TU Bergakademie Freiberg, Freiberg, Germany

Abstract: **The details in the formation of the primary acoustic sources in voice production during phonation are not yet fully understood. Some acoustic sources are due to the unsteady flow evolving between the vocal folds, where a jet develops. The glottal jet flow downstream the vocal folds features characteristics which depend on the physiological conditions, e.g. the parameters in geometry, kinematics, material and fluid. A driven mechanical model of the vocal folds is used with the aim to study the flow details and the acoustic sources in the glottal jet. Numerical simulations and experimental measurements of the flow are carried out. The results show topological characteristics of the glottal jet flow. When prominent ventricular folds are included in the vocal folds model the jet evolves differently due to the interaction with these supraglottal structures. They lead to a changed distribution and abundance of the flow acoustic sources and changed spectral properties of the flow close to the glottis.**

Keywords: **mechanical vocal folds model, ventricular folds, flow simulation, higher harmonics, spectral analysis**

I. INTRODUCTION

The production of the voice is a complex process, which is influenced by a wide range of parameters [1]. In general voicing is a more or less coupled process of fluid-structure-acoustic interaction. The singing and phonation regimes differ quite strongly from each other. Due to its complexity models of the respiratory system, in particular the trachea, the glottis and the vocal tract are generated in order to reduce the problem study to specific voicing aspects. The "pressed" and "breathy" voicing types indicate the importance of the detailed knowledge of the generation of the primary voice source. Singing is a strongly coupled problem and represents a special area in the voicing research with regard to professional singers; whereas phonation is more amenable and has a wider range of application in everyday life. Extensive investigations are nowadays carried out with the aim of tackling voice disorders in phonation. The generation of the primary acoustic signal at the glottis is the first link in the chain of voice production. Herein, the overall output voice signal spectrum is partly influenced by the nature of the flow field downstream of the glottis. Several models exist for the investigation of the primary voice sources: theoretical / lumped mass models, computational fluid dynamics (CFD) models, and mechanical models [2]. These are also classified with regard to the degree of idealization into static, dynamic driven and self-oscillating or 1-dimensional, 2-dimensional / axisymmetric or 3-dimensional (3-D) models. Most of these models incorporate only a very simplified geometry of the vocal tract.

The time-dependent 3-D flow field in a driven vocal folds model [3], which considers the 3-D shape change of the glottis during the cycle of phonation, is considered in the present paper.

The description of 3-D effects in vortex dynamics such as stretching and bending of vortex lines and the determination of the local pressure and velocity are of immense importance for the spectral characterization of the flow field. The temporal evolution of large and small scale flow structures including their interaction with supraglottal walls may change the spectral fingerprint of the flow field. Therefore, flow effects at different driving pressures, changed glottal and supraglottal configurations can be explored with regard to the resulting output flow patterns. In order to combine as best as possible their inherent advantages, experimental and numerical methods of flow investigation are applied simultaneous.

II. METHODS

The main physiological parameters of real vocal folds kinematics during phonation are replicated. The characteristic movement of the walls in the glottal region, e.g. the continuous deformation of the mucosal layer is achieved by means of two 3D contoured cams, which rotate in counter-direction [3]. Similarity of geometry, flow dynamics and fluid dynamic forces is kept in the model. Due to the low Mach number, the flow can be treated as incompressible. The time-dependent and 3D nature of the flow field downstream of the glottis requires flow analysis methods with appropriate temporal and spatial resolutions.

The experimental model of the vocal folds is shown in Fig. 1. The cams which are covered with a membrane can be seen on the left hand side. On the right hand side of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

the photograph supraglottal elements e.g. ventricular folds (VFs) which are optionally inserted into the test section downstream of the vocal folds model are indicated. Two variations of models of the VFs are available: first, rigid transparent models allowing distortion-free optical access into the flow field; second, models with a compliant surface layer and incorporating an air-cushion. The latter are used for pressure sensing the higher harmonics from the glottal jet flow or for selective activation of the compliant surface layer.

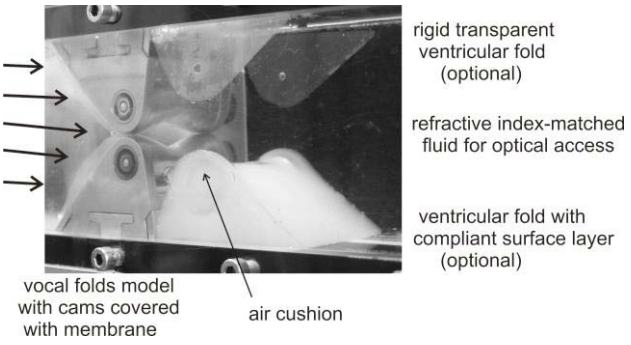


Fig. 1: Photograph of the model of the vocal folds.

In the experiments a global driving pressure head across the glottal orifice is imposed. The close-to-glottis flow-dependent pressure difference is recorded. The well established method of high speed flow visualization is used for accurate determination of temporal and spatial velocity information in selected measuring planes. The resulting glottal volume waveform is measured [3] and given as input for the inlet boundary condition in the numerical model.

The Navier-Stokes equations for incompressible fluid flow are discretized with the Finite Volume method and solved numerically with the open source CFD code OpenFOAM. A block structured mesh of 1 million cells with variation in time according to the glottal kinematics is implemented. The solver uses a second order Crank Nicolson time stepping and as space discretization a second order TVD (total variation diminishing) scheme. The full transient 3-D flow field in the near-glottal region is simulated. The subgrid-scale turbulence is modeled implicitly. All simulations are carried out with the volume waveform synchronous to the imposed time-varying motion of the 3D glottis model.

III. Results

A. *Flow structures*

Fig. 2 shows the experimental visualization of the flow in the mid-coronal plane at the maximum opening instant $t/T_0 = 0.25$ of the glottal cycle. The case without (top) and with rigid (bottom) VFs has been studied. The character of the near field of the emerging glottal jet with its most energetic large coherent vortex structures is shown. Kelvin-Helmholtz instabilities are responsible for the roll-up of the jet edge. The jet front and the successive vortex structures are seen to interact with the VFs. The determination of the pressure fluctuations due to the jet edge instabilities is given in section C.

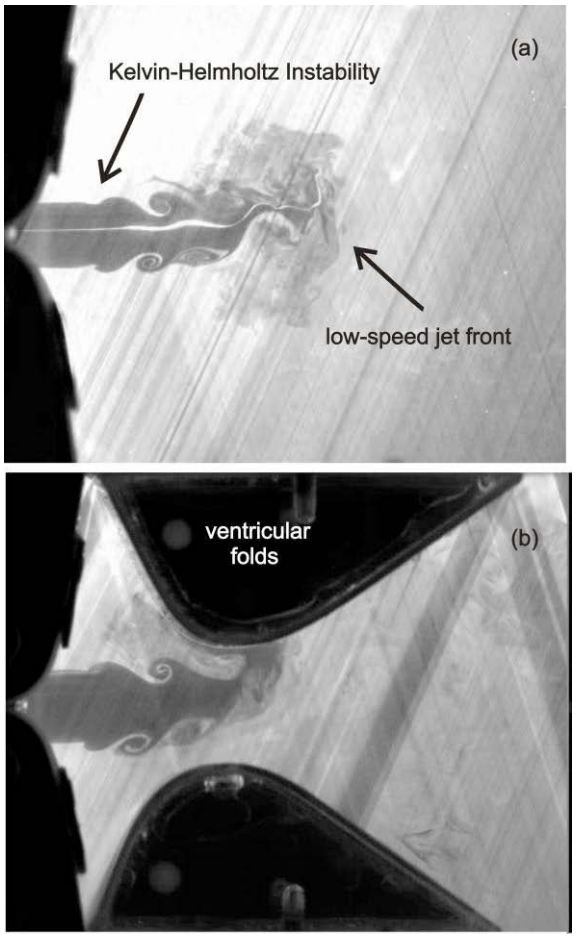


Fig. 2: Visualization of the flow in the mid-coronal plane for transglottal pressure of $\Delta p = 6$ cmH_2O at maximum opening instant $t/T_0 = 0.25$ of the glottal cycle for two supraglottal configurations (a) and (b).

Further flow results are shown from the numerical simulations which resolve the full 3D flow field in space and time in the glottal model. A preliminary study on resolution requirements, accuracy and convergence of the model has been carried out in [4]. There exist several velocity or pressure based tools for vortex detection. In Fig. 3 the Q criterion [5] is used to illustrate the 3-D vortex structures. Elliptic vortex rings are generated at the glottal orifice. These are strongly deformed due to self-induction, interaction among each other and with the supraglottal walls, e.g. VFs.

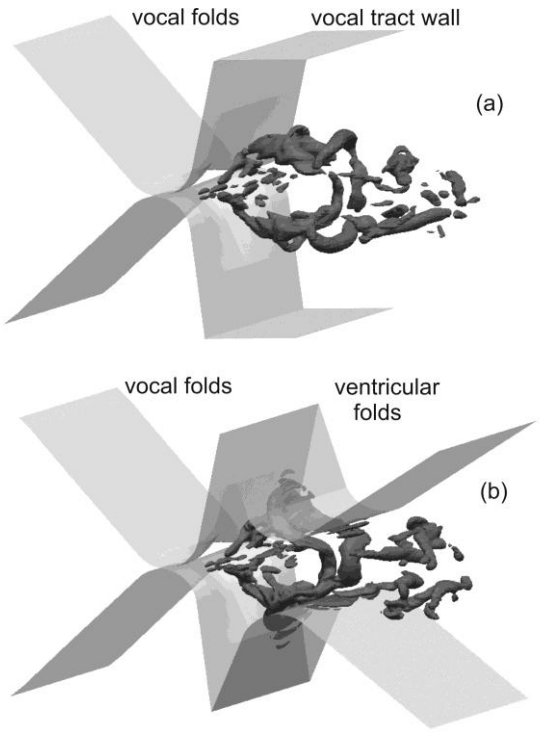


Fig. 3: Representative isocontour of the Q-criterion of the flow field at the divergent closing instant $t/T_0 = 0.35$ of the glottal cycle for both supraglottal configurations (a) and (b) from Fig. 2 at the transglottal pressure of $\Delta p = 6$ cmH_2O.

The complex 3D unsteady vortex structures which are generated downstream of the glottal orifice are already subjected to break-down.

B. Primary acoustic sources

The distribution of the divergence of the Lamb vector **L** can be computed from the velocity **u** of the flow field and appears as a dominant acoustic source term in Lighthill's wave equation [6]. The source term reads

$$\nabla \cdot \mathbf{L} = \nabla \cdot ((\nabla \times \mathbf{u}) \times \mathbf{u}) \quad . \tag{1}$$

One example of this distribution is shown in Fig. 4.

C. Power spectrum

In Fig. 5 the normalized power spectra of the flow field velocity from the numerical simulation is compared for the cases without and with VFs in a center point at a downstream distance from the glottis corresponding to one vocal tract height. The differences are considerable.

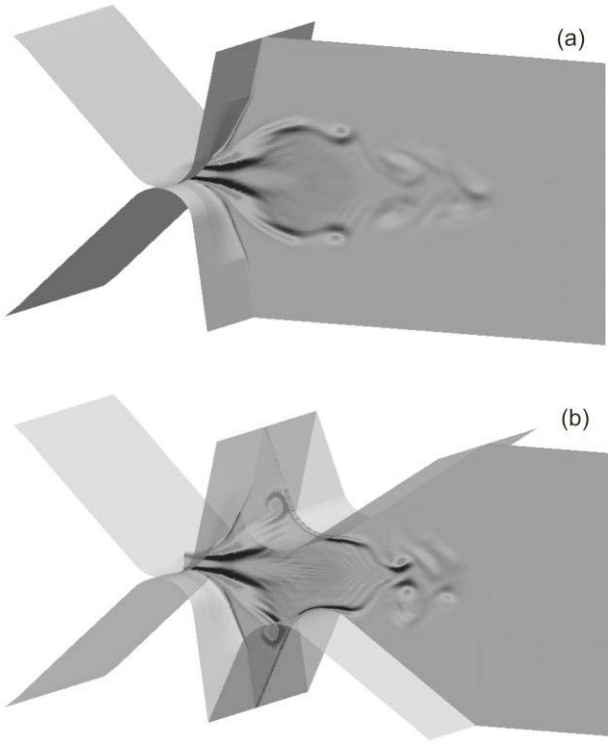


Fig. 4: Distribution of the Lamb vector divergence (black positive, grey 0, white negative value) of the flow field at the divergent closing instant $t/T_0 = 0.35$ of the glottal cycle in the mid-coronal plane for both supraglottal configurations (a) and (b) from Fig. 2.

In order to determine the pressure fluctuations of the jet edge instabilities in experiment, a VF has been replaced with a VF model with a compliant surface layer and air cushion. The small amplitude of the pressure fluctuations poses a challenge in the measurement and analysis of the data. Fig. 6 clarifies the actual situation. The frequency spectra of two pressure measurements are shown in dimensionless form with the help of a Strouhal number Sr defined as

$$\mathrm{Sr} = \mathrm{f} \cdot \mathrm{w} / \mathrm{u}_{\mathrm{mean}} \tag{2}$$

where f is the frequency content of the pressure signal, w is the maximum width of the glottal gap and u_{mean} is the mean velocity in the glottal gap. On top the spectrum results from the reference pressure upstream of the vocal folds model. The spectrum below results from the integral pressure measured in the air-tight air cushion. The difference of both spectra yields the Strouhal number with a value of 0.27, which is supposed to be due to the shear layer instabilities interacting with the walls. This value is highlighted by the arrow in the spectrum and it correlates well with the frequency value from the numerical simulation in Fig. 5.

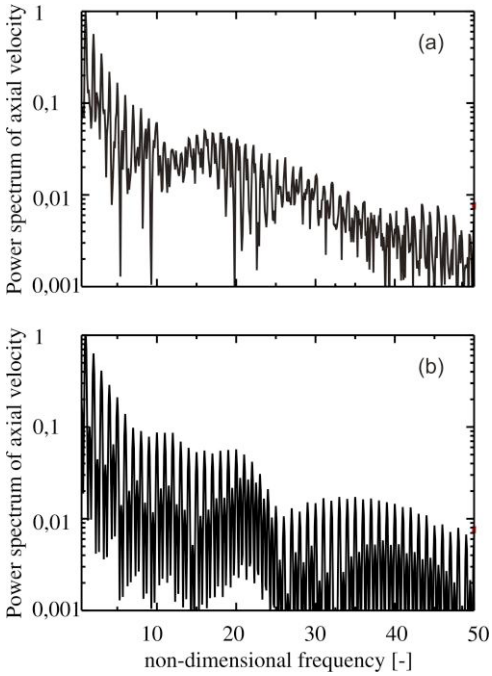


Fig. 5: Power spectrum of axial velocity in a center point at distance from glottis corresponding to one vocal tract height for both supraglottal configurations (a) and (b) from Fig. 2; Non-dimensional frequency related to fundamental frequency.

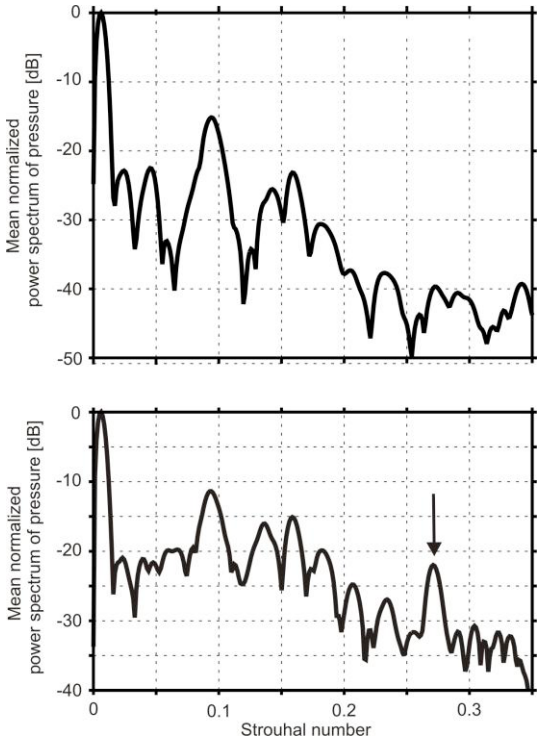


Fig. 6: Mean normalized power spectrum of pressure sensed at upstream position (top) and by VF (bottom).

IV. Discussion

The change in the vortex dynamics and the spectra at different supraglottal configurations are clearly shown. A negative slope of 3 dB per octave in the low frequency range up to the 10^{th} harmonic is extracted the spectra in Fig. 5. Due to the jet edge interaction with the ventricular folds the higher frequency range of the spectrum in configuration (a) differs considerably from that in configuration (b).

V. Conclusion

Prominent ventricular folds leave a strong fingerprint in the spectra of the flow close to the glottis. The ventricular folds redirect part of the displacement flow into the lateral gap of the Morgani space which seems to stabilize the jet core at the exit of the glottis. In addition, the shear-layer roll-up is affected by the presence of the folds and vortex structures are interacting with the walls in this region. As a consequence, vortex dynamics and wall interaction is changed considerably when supraglottal structures are included in the models. These effects are well seen in the change of spectral content ithin the flow. Further studies in our lab now concentrate on possible feedback and jet-control by passive and active excitation of the ventricular folds wall.

References

[1] I.R. Titze, *Principles of Voice Production*, Prentice Hall, New Jersey, 1994.

[2] S. Kniesburges, S. Thomson, A. Barney, M. Triep, P. Sidlof, J. Horácek, Ch. Brücker and S. Becker, "In vitro experimental investigation of voice production," *Current Bioinformatics* (accepted 2011).

[3] M. Triep and Ch. Brücker, "Three-dimensional nature of the glottal jet," *J. Acoust. Soc. Am.* 127 , pp. 1537–1547, 2010.

[4] R. Schwarze, W. Mattheus, J. Klostermann and Ch. Brücker, "Starting jet flows in a three-dimensional channel with larynx-shaped constriction," *Comp. Fluids* 48 , pp. 68–83 , 2011.

[5] J.C.R. Hunt, A.A. Wray and P. Moin, "Eddies, stream, and convergence zones in turbulent flows," *Center for Turbulence Research Rep. CTR-S88*, 1988.

[6] M.S. Howe, "Contributions to the Theory of Aerodynamic Sound, with Application to Excess Jet Noise and the Theory of the Flute," *J. Fluid Mech.* 67 , pp. 597–610, 1975.

SELF-OSCILLATING, MULTI-LAYER NUMERICAL AND ARTIFICIAL VOCAL FOLD MODELS WITH THIN EPITHELIAL AND LOOSE COVER LAYERS

S. L. Thomson, P. R. Murray

Department of Mechanical Engineering, Brigham Young University, Provo, Utah, USA

Abstract: **Synthetic models are used to study vocal fold flow-induced vibration. Advantages include reproducibility and vibration frequencies typical of human phonation. Limitations of recent models include lack of a mucosal wave, excessive inferior-superior motion, and limited convergent-divergent motion. To overcome these limitations, a synthetic vocal fold model was developed that included separate epithelial and lamina propria layers. A corresponding finite element model was developed. High-speed imaging was used to quantify synthetic model motion, including videokymography and determination of three-dimensional marker trajectories. Both models exhibited similar characteristics in terms of vibration frequency (around 115 Hz) and maximum glottal width (just under 2 mm). The synthetic model onset pressure was 0.4 kPa, which is significantly lower than many previous synthetic models. These values are consistent with human phonation. Importantly, in both models mucosal wave-like motion was evident and alternating convergent-divergent intraglottal profiles were seen. These advantages will be useful in future experiments and simulations by providing models that exhibit more life-like response and motion. The two models are described, data are presented, significance of the models is discussed, and suggestions for future work are provided.**

Keywords : **Vocal fold models, artificial models, finite element models**

I. INTRODUCTION

Computational and experimental models are used to study vocal fold flow-induced vibration. Many recent models are based on some variation of the multi-layer structure presented by Hirano [1]. Computational models include high-fidelity Navier-Stokes flow solvers coupled with solid models that include cover, ligament, and body layers [2,3]. Synthetic models include epithelium-lamina propria configurations [4] and two-layer body-cover silicone models [5]. The recent two-layer synthetic models are useful because of their reproducibility, low cost, and ease of parameterization. They have a length scale similar to that of the human vocal folds, have layers with differing stiffness, and vibrate at frequencies typical of human phonation.

Most recent synthetic models are currently limited by three features: unnaturally large inferior-superior displacement, lack of a clear mucosal wave, and higher-than-desired onset pressure (usually 1 to 2 kPa, compared with 0.2 to 0.4 kPa for human phonation). These limitations have been attributed in part to the models' cover layers being stiffer than the human cover. The Young's modulus of elasticity values of the model covers have been around 2 to 3 kPa. By contrast, cover shear modulus values around 0.25 kPa (corresponding to Young's modulus values around 0.75 kPa) at 100 Hz have been measured [6].

To overcome these limitations, a synthetic vocal fold model was developed that includes a cover layer that, as is the case with the human vocal folds, included two distinct layers: a thin epithelial layer and an underlying flexible layer that represented the superficial lamina propria. This synthetic model and a corresponding finite element model are described below. Data are presented which demonstrate significant improvements in terms of model motion and onset pressure.

II. METHODS

A. Synthetic Model

The synthetic model geometry is shown in Fig. 1. Silicone interior layers were fabricated according to the multi-layer rapid prototyping, molding, and casting procedures described in [5,7]. The epithelial layer was created by pouring a silicone mixture over the assembled interior layers. The epithelial layer thickness was estimated to be approximately 0.1 mm. Layer Young's moduli were controlled by varying the pre-cured silicone mixture content; values for the different layers were approximately 11.8 kPa (body), 1.6 kPa (ligament), 0.2 kPa (superficial lamina propria), and 49.8 kPa (epithelium). Tension was applied to a fiber thread that ran anteriorly-posteriorly within the ligament layer to reduce inferior-superior motion. High-speed video imaging (Photron SA3, 3000 frames per second) was used to capture model motion.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

B. Finite Element Model

The finite element model consisted of two-dimensional, fully-coupled fluid and solid domains, as shown in Fig. 2. The solid model incorporated the same geometry as the synthetic model, but with a 50 μm-thick epithelium. The material properties were also the same, with the exception that the superifical lamina propria layer material property was based on a nonlinear stress-strain curve. This curve was governed by the equation

$$\sigma(\varepsilon) = 11.2(e^{10.5\varepsilon} - 1), \quad (1)$$

where σ is stress (Pa) and ε is strain. This yielded a tangent modulus of 200 Pa at $\varepsilon = 0.05$ and 972 Pa at $\varepsilon = 0.2$. The finite element model did not include a fiber.

The fluid model used an incompressible, viscous, 2D, unsteady Navier-Stokes solver with a constant 600 Pa inlet pressure. The solid domain allowed for large strain and large deformation and included Rayleigh damping ($\alpha = 101.67$, $\beta = 0.0001073$) for energy dissipation.

Solution was accomplished using the commercial code ADINA. A time step size of 10^{-4} and a second-order composite time marching scheme were used. For computational efficiency, medial-lateral symmetry was assumed. The fluid domain consisted of 7340/7641 1st-order elements/nodes and the solid domain consisted of 2359/2582 1st-order elements/nodes (see Fig. 3). A solution for 1500 time steps required approximately 3.2 hours on a single 2.53 GHz Intel P9500 processor.

III. Results

The synthetic model had an onset pressure of 400 Pa. At a pressure 20% above onset pressure (480 Pa), the vibration frequency was 114.5 Hz and the maximum glottal width was approximately 1.8 mm. These values compare well with those of human phonation.

Importantly, mucosal wave-like motion was evident and the inferior-superior motion appeared to be lower than with previous two-layer models. To capture this wave-like motion, a hemilarynx configuration and two synchronized high speed cameras (Photron SA3, 3000 frames per second) were used to track the medial surface position in a manner similar to that described in [8]. One sample image is shown in Fig. 4 in which ink dots placed on the model surface are visible. The medial-lateral trajectories (three-dimensional positions) of the dots in the center column were tracked over several oscillation periods, as shown in Fig. 5. A wave-like motion clearly propagated superiorly along the medial surface, and an alternating convergent-divergent medial surface profile was visible. Evidence of this convergent-divergent motion can also be seen in the kymogram shown in Fig. 6 (obtained using a single high-speed camera and a full larynx configuration).

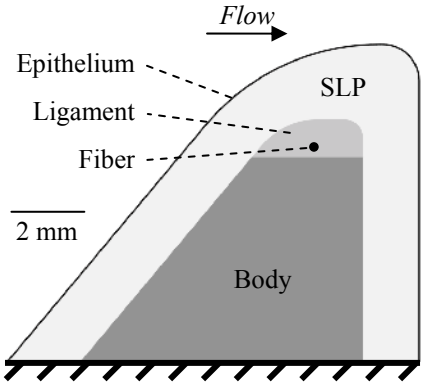


Figure 1. Vocal fold model geometry and length scale.

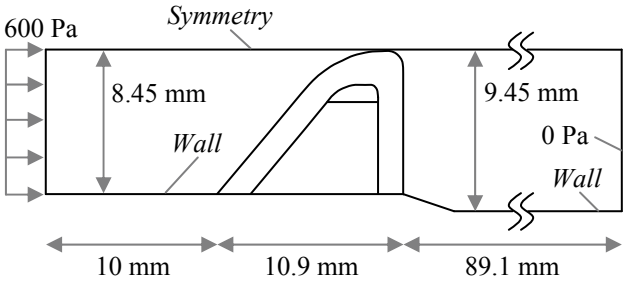


Figure 2. Computational fluid and solid domains.

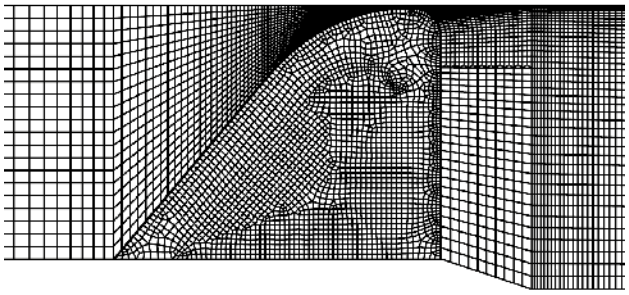


Figure 3. Finite element mesh.

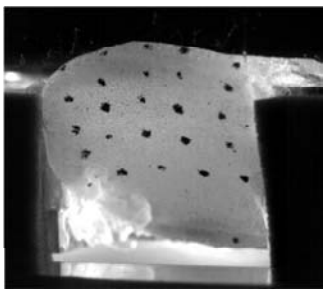


Figure 4. Image of the medial and inferior surfaces of the vocal fold model. Flow is from bottom to top.

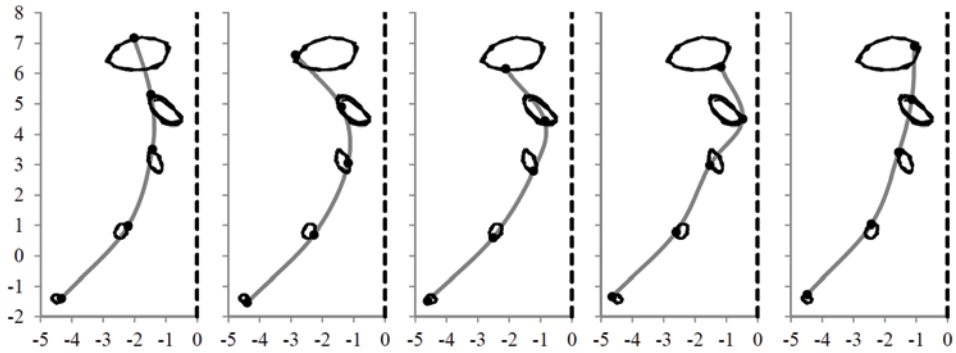


Figure 5. Synthetic model medial surface profile at five instances of one oscillation. Dots denote positions of tracking markers, black lines are tracings of markers over time, and gray lines denote medial surface profiles. Air flow is from bottom to top. Axes denote distance in mm. Solid dashed line denotes location of plate against which synthetic model was vibrating. The pressure was 0.75 kPa and the frequency was 116.2 Hz.

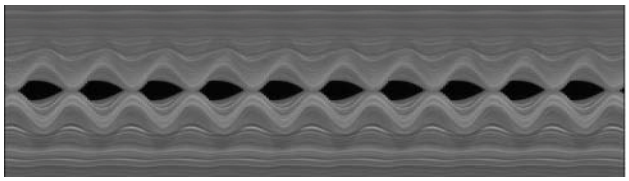


Figure 6. High-speed kymogram of several periods of the synthetic model during flow-induced vibration. The pressure was 0.48 kPa and the frequency was 114.5 Hz.

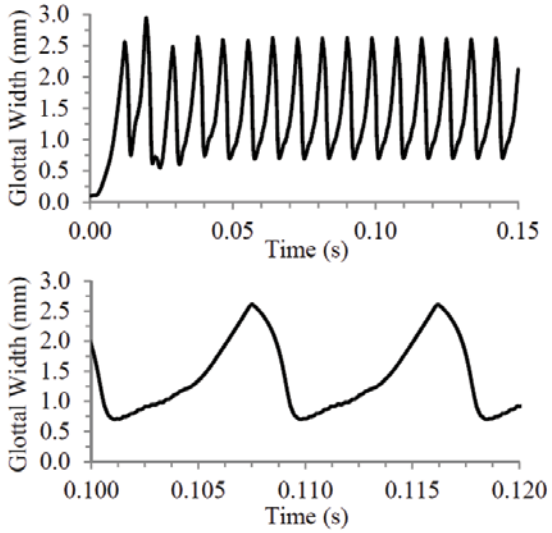


Figure 7. Finite element model glottal width vs. time. Top: entire simulation. Bottom: Two steady-state periods.

The finite element model vibrated at 116 Hz with a maximum glottal width of 1.9 mm, which compares well with the synthetic model response. Glottal width vs. time is shown in Fig. 7. Steady-state vibration was achieved around 0.05 s. Still images of model motion are shown in Fig. 8, in which mucosal wave-like motion is evident.

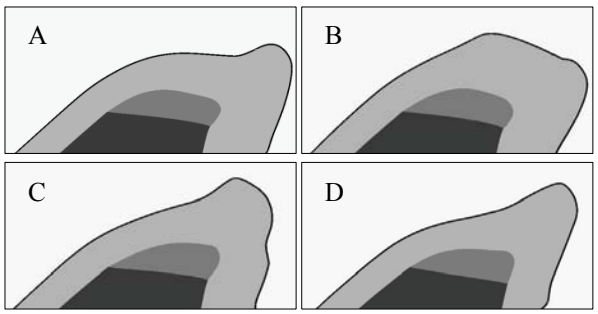


Figure 8. Finite element vocal fold model at four simulation times during one period after reaching steady state. Air flow is left to right.

IV. Discussion

The synthetic and computational models exhibited similar characteristics in terms of vibration frequency and amplitude. Some differences in motion were observed; for example, unlike the synthetic model, the numerical model did not experience complete glottal closure during vibration. Differences in motion of the two models were attributed to four factors: differences in material properties (stress vs. strain relationships, Poisson's ratio, and damping coefficients), three-dimensionality of the synthetic model versus two-dimensionality of the finite element model, difference in thickness of the epithelial layer, and presence of an anterior-posterior fiber in synthetic model.

V. Conclusion

Complementary synthetic and finite element models of the vocal folds have been developed and tested. The models were based on the same multi-layer geometry. Each included a cover layer that was comprised of a thin epithelial layer and a very flexible layer that was similar to the superficial lamina propria. Each also included

ligament and body layers, and the synthetic model included a fiber imbedded within the ligament layer.

In both models mucosal wave-like motion was evident. Alternating convergent-divergent intraglottal profiles were also seen. The vibration frequencies and glottal amplitudes were typical of adult human male phonation. Further, the synthetic model had an onset pressure that was much lower than previous models and that is comparable to human phonation. These advantages will be useful in future experiments and simulations by providing models that exhibit more life-like response and motion.

For both models future work includes the use of anisotropic materials. Incorporation of a downstream duct (to simulate the vocal tract) in the synthetic model will also be important. For the finite element model, future work includes performing extensive numerical verification studies (e.g., ensuring that the solutions are independent of grid density and time step size), extending the model to three dimensions, and removing the symmetry condition. The latter will enable the study of asymmetric aerodynamics and vocal fold properties. Potential future work also includes investigation of the influence of epithelial layer thickness and of the material properties of the different layers on model response.

VI. Acknowledgements

This study was supported by Grants R01DC9616 and R01DC5788 from the U.S. National Institute on Deafness and Other Communication Disorders (NIDCD). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCD or the National Institutes of Health.

References

[1] M. Hirano, "Structure and vibratory behavior of the vocal folds," in *Dynamic Aspects of Speech Production*, M. Sawashima and F.S. Cooper, Eds. Tokyo, Japan: University of Tokyo Press, 1977, pp. 13-30.

[2] F. Alipour, D.A. Berry and I.R. Titze, "A finite-element model of vocal-fold vibration," *J. Acoust. Soc. Am.*, vol. 108, pp. 3003-3012, 2000.

[3] X. Zheng, S. Bielamowicz, H. Luo and R. Mittal, "A computational study of the effect of false vocal folds on glottal flow and vocal fold vibration during phonation," *Ann. Biomed. Eng.,* vol. 37, pp. 625-642, 2009.

[4] R.W. Chan, I.R. Titze and M.R. Titze, "Further studies of phonation threshold pressure in a physical model of the vocal fold mucosa," *J. Acoust. Soc. Am.,* vol. 101, pp. 3722-3727, 1997.

[5] T. Riede, I.T. Tokuda, J.B. Munger and S.L. Thomson, "Mammalian laryngeal air sacs add variability to the vocal tract impedance: Physical and computational modeling," *J. Acoust. Soc. Am.*, vol. 124, pp. 634-647, 2008.

[6] R.W. Chan and M.L. Rodriguez, "A simple-shear rheometer for linear viscoelastic characterization of vocal fold tissues at phonatory frequencies," *J. Acoust. Soc. Am.*, vol. 124, pp. 1207-1219, 2008.

[7] P.R. Murray and S.L. Thomson, "Synthetic, multi-layer, self-oscillating vocal fold model fabrication," *J. Vis. Exp.* (accepted).

[8] M. Doellinger and D.A. Berry, "Visualization and quanitification of the medial surface dynamics of an excised human vocal fold during phonation," *J. Voice*, vol. 20, pp. 401-413, 2006.

Session III:
Signal analysis

AUTOMATED TOOLS FOR IDENTIFYING SYLLABIC LANDMARK CLUSTERS THAT REFLECT CHANGES IN ARTICULATION

Suzanne Boyce[1], Harriet Fell[2], Lorin Wilde[3], and Joel MacAuslan[3]

[1]University of Cincinnati, OH, [2]Northeastern University, Boston, MA,
[3]Speech Technology and Applied Research, Bedford MA, USA

Abstract: **We have developed a set of software tools to detect articulatory changes in the production of syllabic units based on acoustic landmark detection and classification. Results from the application of this automatic analysis system to studies of Parkinson's Disease and Sleep Deprivation show the ability to detect subtle change. We are making these tools available as add-ons to systems such as Wavesurfer and R.**

Keywords: **speech-acoustic landmarks, syllabic landmark cluster, automatic vocalization processing.**

I. INTRODUCTION

Acoustic evidence provides information on speech production, but that information is scattered across multiple frequency bands and multiple time scales. Landmark analysis [5,6] is one approach by which acoustic patterns characteristic of particular changes in speech movements are detected. In this paper, we describe an extension of the landmark method to the detection of articulatory complexity in the production of syllables, by using clusters of landmarks as a measure of whether a string of (intended) syllables is produced in its canonical form (dictionary pronunciation), in a less complex (CCVC -> CVC), or more *lenited* form (softened consonants). We refer to this measure as a measure of syllabic complexity, and to our landmark cluster measure as a "syllabic cluster" measure. We have applied this approach successfully to measure speech articulation changes in Parkinson's Disease, in infant speech development, in sleep deprivation, and other studies.

The notion of syllabic complexity is illustrated as follows. A word such as "interesting" can have four syllables in its canonical form, but when uttered as /ɪnrɛstɪŋ/ it has three syllables with fewer consonants, and thus reduced articulatory complexity. In landmark systems in general, different types of types and combinations of speech sounds are detected as different patterns of landmarks. In our particular system, a syllabic landmark cluster is a sequence of consecutive landmarks grouped according to specific rules. For example:

1. A syllabic cluster must contain a voiced region of at least 30 ms, corresponding to a syllable nucleus.

2. A noisy sound such as "s" (/s/) must hit a threshold of loudness before being detected.

If uttered in a canonical fashion, the pronunciation of a word will show a characteristic pattern of landmarks for each syllable in that word. As long as the syllables are uttered with the same acoustical characteristics, our measures will detect the same pattern of landmarks. However, if the syllables are uttered less canonically—perhaps with less extreme articulatory movements, less precise timing, or reduced aerodynamic support—--- then fewer landmarks will be detected. Our version of the speech-acoustic landmark system thus can be used to detect two common effects in speech production: (1) simplification of syllable onsets (e.g. "string" /strɪŋ/ as /srɪŋ/), nuclei (e.g. "diamond" /dɑɪmənd as /dɑmənd/) and rimes (e.g. "pelt" /pɛlt/ as /pɛl/, and (2) fewer uttered syllables.

II. METHODS: LANDMARK SYSTEM

Landmarks and Rules: Our landmark analysis system is based on Stevens *et al.* [6], especially as developed by Liu [5] and Howitt [4]. The speech signal is automatically partitioned into 5 frequency bands plus a voicing-status contour. Abrupt landmarks are identified as points where abrupt changes in the amplitude of several frequency bands coincide in a specified pattern [5,6]. These landmark patterns are identified by comparison between "coarse" and "fine" temporal resolution.

The system detects the following types of landmarks:

1. g: glottis. Marks a time point at which voicing begins (+g) or ends (-g), based on the harmonic spectrum.
2. s: syllabicity. Marks sonorant consonantal releases (+s) and closures (-s).
3. b: burst. Marks frication onsets or affricate/stop bursts (+b) and points where aspiration or frication ends (–b) due to a stop closure.
4. V: vowel. Marks a time point corresponding to maximum harmonic power.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

The +/- b and +/- s "abrupt" landmarks are identified from patterns of rapid change in the amplitude of several frequency bands. The +/-g and V landmarks are identified from the harmonic spectrum.

This system makes no attempt to identify phonemes, but it is sensitive to broad categories of speech sounds and to aspects of metrical structure. The features it detects are those known as "articulator free" [6] because they are independent of the specific articulator used to produce the segment. These features are instead associated with creation and release of constrictions in the vocal tract and with the acoustic consequences of those constrictions and releases.

An example of how abrupt landmarks are determined from patterns across frequency and voicing bands is shown in Fig. 1. An example of landmark location in the speech signal can be found in Fig. 2, which shows a spectrogram of the nonsense word /pʌtəkə/ repeated 10 times in two breath groups by a native speaker of American English with moderate dysarthria due to Parkinson's Disease.

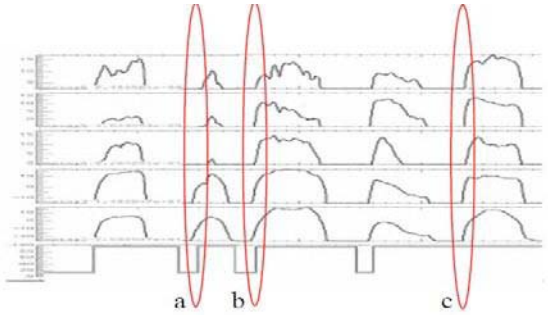


Figure 1. Spectral analysis of an utterance: voicing (bottom) and five frequency bands' energy waveforms. (a) Too few bands show large, simultaneous changes in energy. (b) All bands show large, simultaneous energy increases immediately before the onset of voicing, identifying a +b (burst) landmark. (c) All bands show large, simultaneous energy increases during ongoing voicing, identifying a +s (syllabic) landmark.

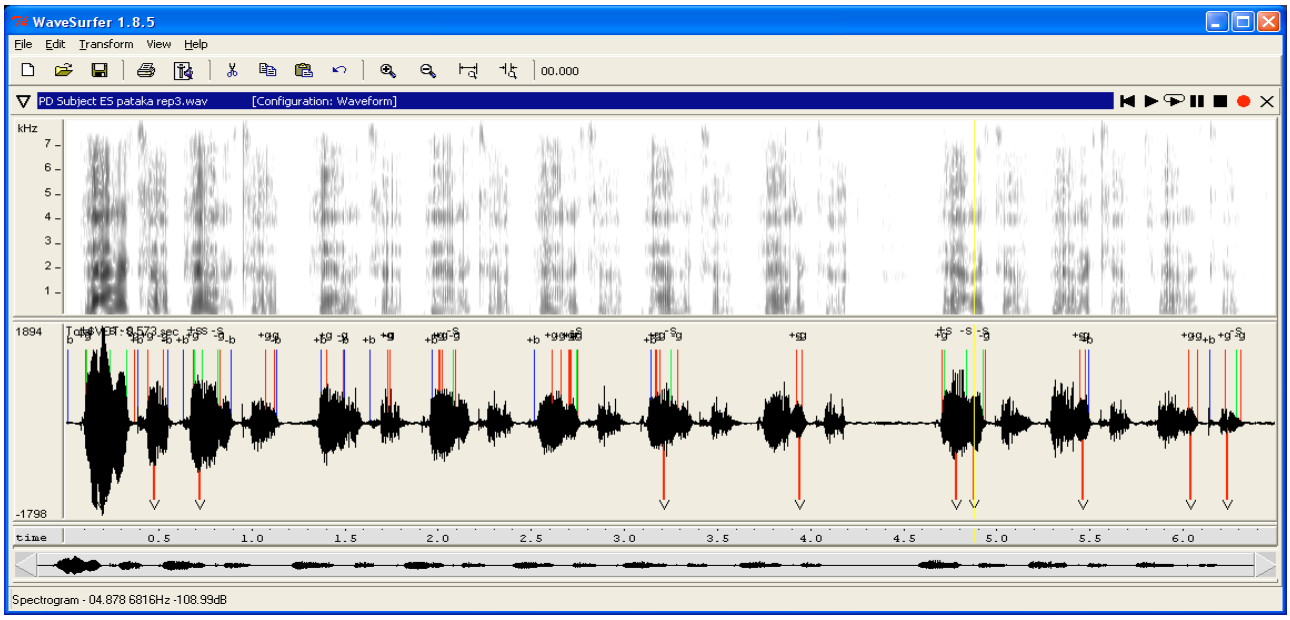


Figure 2. Ten repetitions of /pʌtəkə/ by an American English speaker with moderate dysarthria due to Parkinson's Disease. Vertical lines above the waveform pane show +/- b, +/-s and +/- g landmarks. Vertical lines below the waveform pane show Vowel landmarks as V. The period of silence shows the pause between breath groups.

Use of the Landmark System to Characterize Differences in Speech Production: The landmark system operates with empirically derived threshold values. As discussed, abrupt landmarks are determined by the patterns of abrupt change across frequency and voicing bands; if the amplitude value of the signal in a particular set of frequency and voicing bands meets the predetermined threshold for abruptness, then a landmark is detected. If the amplitude value of the signal in any of the frequency/voicing bands does not meet this criterion, then no landmark is detected.

The operation of this system is shown by the pattern of V landmarks in Fig. 2. As noted above, the speaker produced /pʌtəkə/ in two breath groups; the first seven repetitions belong to the first breath group, and the

following three repetitions belong to the second breath group. This speaker showed a tendency to dysphonia typical of Parkinson's patients, characterized subjectively as causing a harsh and breathy voice, and the dysphonic phonation was more marked in the late portions of a breath group—presumably because reduced breath support made it more difficult to sustain normal periodic vocal fold vibration. Because the V landmarks are computed on the basis of harmonic power, and dysphonic vowels are produced with less harmonic power, fewer V landmarks will be detected on dysphonic voices. This effect is shown in Fig. 2, where the first few repetitions in the first breath group are marked with V landmarks on the stressed syllable, while the last few repetitions in the same breath group show that no such landmarks have been detected. Note that these repetitions were produced with vowels—this is evident in the spectrogram--but the vowels had too little harmonic power to be registered as V landmarks.

Grouping Landmarks to Characterize "Syllabic Clusters": Fell & MacAuslan originally developed the syllabic cluster measure to detect the increasing syllabic complexity of utterances by young children [2, 3]. More recently, we have applied this method, termed the Syllabic Cluster analysis, to speech uttered under normal and sleep-deprived conditions, and to speech by Parkinson's Disease patients undergoing Deep Brain Stimulation (DBS) therapy.

Cluster Rules: The Syllabic Cluster analysis works by grouping sequences of detected landmarks into clusters that roughly correspond to syllabic units in the acoustic speech signal. The grouping rules include categorical dependencies as well as dependencies of timing, and were empirically determined from datasets of speech.

For example, one such rule states that a gap of 30 ms in voicing, with whatever ±b's immediately follow it, identifies a type of syllable cluster endpoint. In contrast, burst-like noise that does not occur within 120 ms before a voiced region, or 80 ms after, is not part of a cluster. Indeed, we have found it useful to designate these types of isolated bursts as non-speech noise. The syllabic grouping procedures are described in more detail in Fell et al. [2,3] and Boyce et al. [1] The following is a list of examples of some common types of syllabic cluster that occur in speech:

- (+g,-g)- singleton V [vowel] or CV [consonant-vowel] syllables, where C is voiced;
- (+g,-s) - V or voiced-CV syllables followed by a sonorant consonant and syllabic cluster;
- (+s, -g) - V or voiced-CV syllables, preceded by a syllabic cluster;
- (+g,-s,-g) - VS syllable, where S is a sonorant consonant or voiced obstruent adjacent to the +g or -g;
- (+b,+g,-g) - syllable beginning with fricative: (+b) marks the presence of frication;
- (+b,-b,+g,-g) - syllables with an initial plosives: (+b , -b) mark the beginning and end of the release.

III. METHODS: APPLICATION

Parkinson's Disease Study: In one study using the Syllabic Cluster measure, we contrasted speech as produced by Parkinson's Disease (PD) patients who were receiving Deep Brain Stimulation (DBS). In the typical progression of Parkinson's Disease, patients show clinically significant levels of unintelligible speech later than they show gross motor symptoms. Thus, patients in DBS programs may not be showing clinically overt signs of dysarthric speech. However, the application of DBS therapy can sometimes cause their speech intelligibility to worsen, and this is both a matter of clinical concern and scientific interest. The data described in Fig. 3 come from a study of 12 Control vs 15 PD patients who had undergone surgery for Deep Brain Stimulation (DBS) repeating the syllable /ka/. The aim of the study was to detect subtle and/or overt changes in speech production when DBS stimulus was OFF vs. ON.

Sleep Deprivation: In another study, we used the Syllabic Cluster analysis to test whether speech articulation changes as a result of sleep deprivation. Studies of both speech articulation per se, and listener perceptions of change, have shown conflicting results to date [1]. In our study, the speech of 17 speakers of American English (9 female, 8 male) was recorded at 8 hour intervals over 32-40 hours without sleep. (Not all subjects completed the final session.) Subjects read aloud the Rainbow Passage each time. To control for the possible effect of familiarity with the speech materials, another set of 15 subjects (7 male and 8 female) read aloud the Rainbow Passage at 8-hour intervals while maintaining a normal sleep schedule.

III. RESULTS AND DISCUSSION

Parkinson's Disease Study: The mean cluster rate in rapid repetitions of the syllable /kɑ/ decreases (a) between Control vs. PD speakers, and (b) as a result of DBS. The differences were significant at the .01 level.

Sleep Deprivation: The first two sessions were combined as the Early, or Rested, condition. The last two sessions were combined as the Late, or Sleep Deprived condition. As Fig. 4 shows, Syllabic Cluster rate decreased between the Early and Late sessions. This difference was significant at the $p < .05$ level by a binomial (sign) test. In contrast, the Early vs. Late sessions were not significantly different for speakers performing the identical task while following their normal sleep schedule ($p > .10$ by a binomial (sign) test.

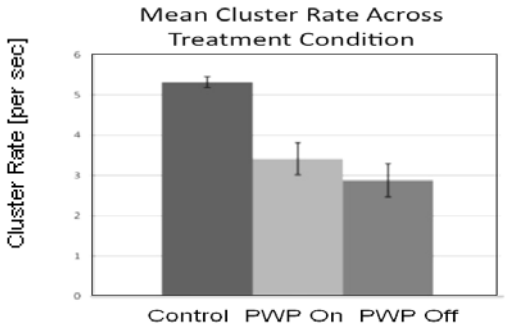


Figure 3. The mean rate of Syllabic Cluster occurrence for 12 age and gender-matched control subjects (Control) vs 15 speakers of American English with Parkinson's Disease (PWP) across Stimulus ON and Stimulus OFF conditions.

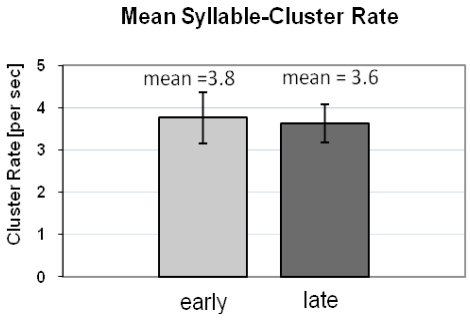


Figure 4. The mean rate of Syllabic Cluster occurrence for 17 speakers of American English reading the Rainbow Passage aloud in Early vs. Late sessions of a 30-40 hour period without sleep.

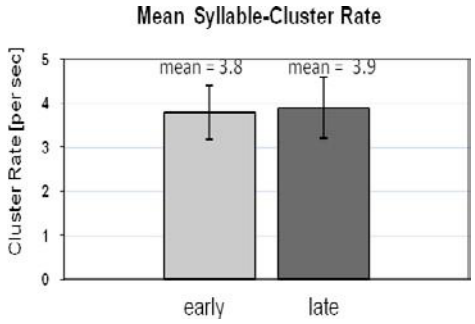


Figure 5. The mean rate of Syllabic Cluster occurrence for 15 speakers who read the Rainbow Passage aloud in Early vs. Late sessions while following their normal sleep patterns.

VI. CONCLUSION

The Syllabic Cluster analysis based on acoustic landmark detection appears to be sensitive to articulatory differences in speech production scattered across multiple frequency bands and multiple time scales. The Parkinson's Disease results suggest this analysis provides a rough measure of a speaker's ability to repeat speech materials with a certain level of articulatory precision at a particular speech rate. The Sleep Deprivation results suggest that speech articulation does indeed change with sleep deficit in a way that reduces the rate at which well-formed syllabic clusters are produced and that this change is not due to familiarity with the speech materials. Both sets of results suggest the analysis is sensitive to very subtle changes that listeners do not always detect. The automatic nature of the analysis facilitates evaluation of large amounts of data.

We are currently developing a set of software tools for automatic landmark detection, and classification into syllabic cluster patterns, to be available as add-ons to systems such as Wavesurfer and R.

ACKNOWLEDGEMENT

Support: R43 DC010104, R42 AG033523, R21 HL086689, R42 HD34686 from the U. S. National Institutes of Health.

REFERENCES

[1] S. Boyce and J. MacAuslan, "Effects of sleep deprivation on speech articulation and intelligibility in noise," *Proc. Of ASMA* (Aerospace Medical Association), Anchorage, AK. 2011.

[2] H. J. Fell, J. MacAuslan, L. J. Ferrier, K. Chenausky, "Automatic babble recognition for early detection of speech related disorders," *Journal of Behaviour and Information Technology*, 1999, 18, no. 1, 56-63.

[3] H. J. Fell, J. MacAuslan, L. J. Ferrier, S. Worst, & K. Chenausky, "Vocalization age as a clinical tool," *Proc. of ICSLP* (International Conference on Speech Processing), Denver, USA, 2002.

[4] A. W. Howitt, *Automatic syllable detection for vowel landmarks*, doctoral thesis M.I.T., Cambridge, MA. 2000.

[5] S. Liu, S. *Landmark detection in distinctive feature-based speech recognition*, doctoral thesis M.I.T., Cambridge, MA. 1995.

[6] K. N. Stevens, S. Manuel, S. Shattuck-Hufnegel, and S. Liu, "Implementation of a model for lexical access based on features," *Proc. Int. Conf. Spoken Language Processing*, Banff, Alberta, 1, 499-502. 1992.

SPECTRAL ANALYSIS OF PATHOLOGICAL VOICES: SUSTAINED VOWELS vs RUNNING SPEECH

R. Fraile[1], J.I. Godino-Llorente[2], N. Sáenz-Lechón[2], J.M. Gutiérrez-Arriola[2], V. Osma-Ruiz[2]

[1]Universidad CEU Cardenal Herrera, Moncada, Valencia, Spain
[2]Department of Circuits & Systems Engineering, Universidad Politécnica de Madrid, Madrid, Spain

Abstract: **A short-time spectral analysis of normophonic and dysphonic voices is presented. This analysis has been performed on recordings of both sustained vowels and running speech for comparison purposes. The reported results indicate that pathological voices tend to have a higher concentration of energy in the lowest frequencies (20 to 300 Hz) and less energy in frequencies from 630 to 1,270 Hz. Additionally, pathological voices tend to experience quicker variations in spectral energy from 770 to 1,720 Hz.**

Keywords: Speech analysis, spectral analysis, correlation

I. INTRODUCTION

The acoustic analysis of voice for clinical purposes has traditionally been made on sustained vowels [1] and a set of parameters measuring voice instability are of common use in clinical software for voice analysis [2]. However, extrapolating the use of such measures to running speech seldom provides good results [3], since the stationarity assumption only holds for sustained phonations. As for spectral domain, the lack of stationarity can be handled by means of short-time spectral measures. Thus, short-time spectral analysis may be useful for analyzing voice quality in running speech [4]. Within this paper, a short-time spectral analysis of normophonic and dysphonic voices is presented. This analysis has been performed on recordings of both sustained vowels and running speech for comparison purposes.

For the spectral analysis, the speech segments of the processed recordings have been passed through a filter bank so as to split the signal into the 22 first critical bands of the human auditory system, as identified by Zwicker [5]. For each band, the instantaneous energy has been computed and, subsequently, the autocorrelation function of each band energy sequence has been calculated. Both the absolute values of instantaneous energy and the width of the autocorrelation functions have been analysed. The reported results allow identifying relevant differences in spectral energy between normophonic and dysphonic voices. In addition, the width of the autocorrelation function is used as a cue for spectral stability. Results indicate that dysphonic voices tend to be more unstable than normophonic voices but only in bands above the 8th one (over 770 Hz)

II. MATERIALS

Processed voice recordings were taken from the Voice Disorders Database distributed by Kay Elemetrics [6]. Specifically, a subset of 53 normophonic and 173 dysphonic voices was selected [7]. For each voice two recordings were available: one corresponding to a sustained phonation of the vowel /æ/ and another corresponding to running speech, namely a fragment of the rainbow passage: "*When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon*". For all recordings, the sampling frequency was 25 kHz and they were normalized in amplitude to have unit mean square value.

III. METHODS

Speech detection was performed on recordings corresponding to running speech. Specifically, a simple detector based on short-time energy and short-time zero-crossing rate was used [8]. Subsequently, both sustained vowel recordings and speech segments in running speech recordings underwent the same process. The first step consisted in the following time-frequency representation.

The short-time spectrogram of a discrete-time signal can be written as:

$$S_p(k) = \sum_{n=-N+pL}^{N+pL} x[n]\,w[n-pL]\,e^{\frac{-j2\pi k(n-pL)}{N_{\mathrm{DFT}}}} \qquad (1)$$

where p is the frame number, k / N_{DFT} is the normalized frequency, L is the number of samples between consecutive frames, $w[n]$ is the framing window which has a length equal to $2N+1$ and $N_{\mathrm{DFT}} \geq 2N+1$ is the number of points of the discrete Fourier transform (DFT).

Defining the framing window to be symmetric and if $w[n] = 0 \;\; \forall |n| > N$, then (1) can be written as follows:

$$S_p(k) = \sum_{n=-\infty}^{\infty} x[n]\,w[pL-n]\,e^{\frac{j2\pi k(pL-n)}{N_{\mathrm{DFT}}}} \qquad (2)$$

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Since (2) has the form of a convolution, the sequence of values of the spectrogram corresponding to the k^{th} frequency can be written as a convolution followed by decimation by a factor L:

$$S_p(k) = S_k[p] = \left(x[n] * h_k[n]\right)_{n=pL} \quad (3)$$

being

$$h_k[n] = w[n]e^{\frac{j2\pi kn}{N_{DFT}}} = w[n]e^{j\Omega_k n} \quad (4)$$

While the usual spectrogram has a common window $w[n]$ for all values of k, the formulation in (3) allows defining different windows for different frequency bands, hence $w_k[n]$ instead of $w[n]$.

For the herein reported work, $w_k[n]$ were chosen to be hamming windows with odd length. The specific length of each one was selected so that its -3 dB bandwidth matched the width of the k^{th} critical band. The odd lengths allowed integer group delays that permitted subsequent time alignment of all the 22 resulting sequences. Also, each value of Ω_k was selected such that $\Omega_k \cdot f_s / 2\pi$ was equal to the center frequency of the k^{th} band. Fig. 1 provides an overview of the whole scheme.

In a second step, the instantaneous energy, i.e. square modulus, of every filter's output was calculated: $e_i[n] = |x_i[n]|^2$. Last, the normalized autocorrelation function of each energy signal was computed.

$$\rho_i[m] = \frac{E\left\{\left(e_i[n] - E\{e_i[n]\}\right)\left(e_i[n+m] - E\{e_i[n]\}\right)\right\}}{E\left\{\left(e_i[n] - E\{e_i[n]\}\right)^2\right\}} \quad (5)$$

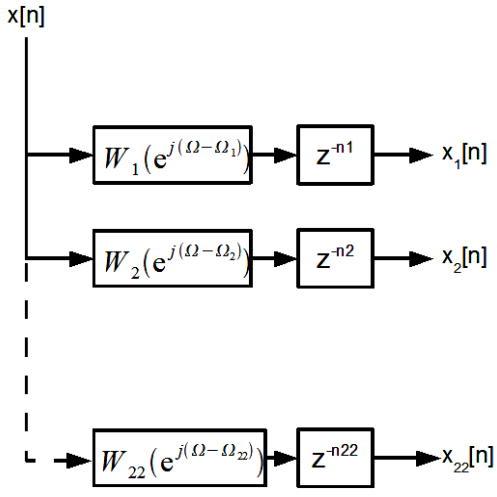


Figure 1. Filterbank that splits the signal into the 22 first critical bands.

After (5), the band energy decorrelation time was defined as the highest time shift $m \cdot f_s$, being f_s the sampling frequency, for which $\rho_i[m] \geq 0.5$. The concept of decorrelation time is illustrated in Fig. 2.

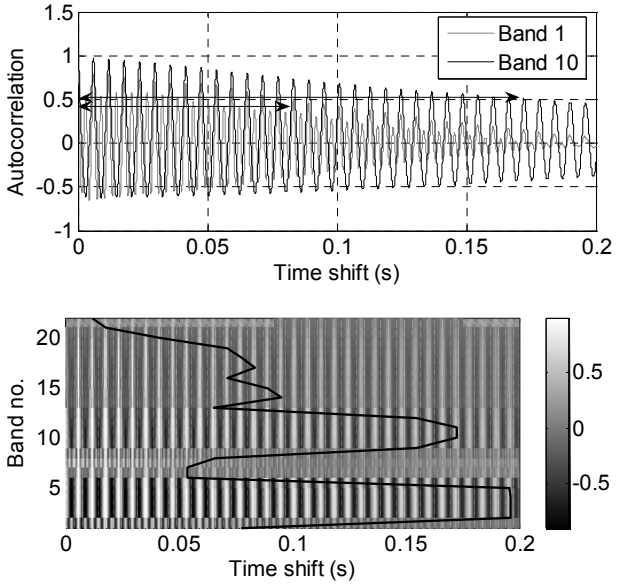


Figure 2. *Top*: Autocorrelation of the energy signals corresponding to two bands (critical bands 1 and 10). Both decorrelation times have been indicated with double arrows: the longest corresponds to band 10 and the shortest to band 1. *Bottom*: Autocorrelation of all 22 band energy signals plotted as gray levels. The continuous line indicates the dependence of decorrelation time on band number.

IV. Results

For both sustained vowels and running speech, the instantaneous band energies $e_i[n]$ were averaged to yield 22 mean band energies per voice record. In the case of sustained vowels, averaging was performed along the full record lengths while for running speech averaging was carried out only along speech segments. Fig. 3 shows the median and the 25th and 75th percentiles of the mean band energies for both sustained vowels and speech segments in running speech. In dysphonic voices, there is a significantly larger portion of energy distributed in critical bands 1 to 4 with respect to the case of normal voices. In the case of sustained vowels, this feature corresponds to a notably lower amount of energy in bands 8 to 10 (770 to 1,270 Hz). In running speech, the difference in energy along the first bands is lower and the spectral range for which dysphonic voices have less energy spans from the 7th to the 14th critical band (630 to 2,320 Hz).

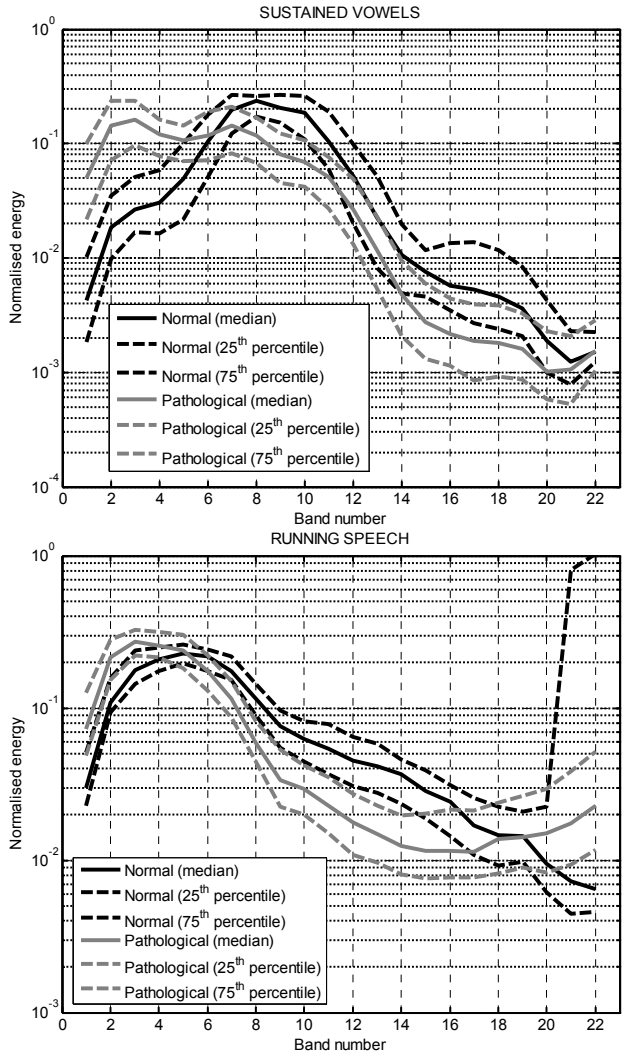


Figure 3. Energy distribution along the 23 first critical bands for sustained vowels (top) and speech segments in running speech (bottom). Continuous lines represent the median value among all recordings, in gray for pathological voices and black for normal voices. Dashed lines indicate the 25th and 75th percentiles.

Within Fig. 4, the median and 25th and 75th percentiles of the band energy decorrelation times are depicted. In running speech, the expectation operator $E\{\cdot\}$ in (5) has been applied only to speech segments. Shorter decorrelation times indicate faster decays in the autocorrelation function and this, in turn, is a cue of quicker variations in the characteristics of the signal ($e_i[n]$ in this case). As shown by the graphs, pathological voices tend to have energy variations in high frequency bands significantly faster than normal voices. This difference seems to be relevant in bands above the 15th one, i.e. frequencies above 2,500 Hz, for the case of sustained vowels. However, such trend does not occur equally in the case of running speech. In this case, results indicate a slighter trend of pathological voices to exhibit shorter decorrelation times in bands 8 to 12 (770 to 1,720 Hz) but the distributions of decorrelation times almost completely overlap for bands beyond the 16th (3,150 Hz).

V. Discussion

The reported results indicate that pathological voices tend to have a higher concentration of energy in the lower critical bands (1 to 3). Such fact, especially in what affects the first band (20 to 100 Hz), is related to the lack of periodicity of the voice. In fact, for fundamental frequencies above 100 Hz, the energy corresponding to the first band is only related to inter-period variations. In the case of running speech, articulation is one evident cause for aperiodicity and this is reflected by normal voices having more energy in the lowest frequency bands than in the case of sustained vowels. Typical phoneme durations around 100 ms [9][10] are related to frequency components in the range of hertzs or tens of hertzs, that is, frequencies corresponding to the lowest bands.

In contrast, pathological voices have similar low-frequency energies in both cases (sustained vowels and running speech); thus, for pathological voices the main cause for lack of stationarity does not seem to be articulation, but pathology itself.

Considering these results, another common feature of both kinds of phonations, which should allow distinguishing between normal and pathological voices, is the ratio of low to high frequency energy; hence a measure of spectral tilt. Spectral tilt has been reported to be a good indicator of breathiness [11]. Herein described results indicate that spectral tilt, measured as the ratio of energy in bands 1 to 3 (20 to 300 Hz) to energy in bands 7 to 10 (630 to 1,270 Hz), should also be a good indicator of dysphonia, both for sustained vowels and for running speech. The use of linear-phase filters in the filterbank of Fig.1 allows such ratio to be computed in short term, at rates up to f_s.

As for the band energy decorrelation time, a measure of band energy variability, while for sustained vowels this measure provides a fair distinction between normal and pathological voices for the highest frequency bands, this is not the case for running speech. Only a minor discrimination ability is to be expected for bands 8 to 12 (770 to 1,720 Hz). Yet, from the absolute value of the decorrelation time, a relevant indication can be obtained for the design of short-term spectral processing schemes: frame sequences with a rate below 100 frames/second (10^{-2} seconds/frame) would not be able to adequately capture the energy variability of pathological voices in bands above the 16th (3,150 Hz) for sustained vowels or, alternatively, the 13th (2,000 Hz) for running speech. Furthermore, for higher frequency bands a frame rate of

at least 1,000 frames per second seems to be required, due to decorrelation times in the order of 1 ms.

ACKNOWLEDGEMENTS

This research work has been financed by the Spanish government through the project grant TEC2009-14123-C04-02. It has also been realized within the framework of COST Action 2103.

REFERENCES

[1] V. Parsa, and D.G. Jamieson, "Acoustic Discrimination of Pathological Voice: Sustained Vowels Versus Continuous Speech" *J Speech Lang Hear Res* vol.44, pp.327-339, 2001.

[2] D.D. Deliyski, "Acoustic Model and Evaluation of Pathological Voice Production", in *Eurospeech* 1993, pp.1969-1972.

[3] Y. Zhang, J.J. Jiang, "Acoustic Analyses of Sustained and Running Voices from Patients with Laryngeal Pathologies", *J Voice*, vol.22, pp.1-9, 2008.

[4] K. Umapathy, S. Krishnan, V. Parsa, and D.G. Jamieson, "Discrimination of Pathological Voices Using a Time-Frequency Approach", *IEEE Tran Biomed Eng*, vol.52, n. 3, pp.421-430, 2005.

[5] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)", *J Acoust. Soc. America*, vol. 33, n.2, p. 248, 1961.

[6] Massachusetts Eye and Ear Infirmary. *Voice Disorders Database*. 1994.

[7] V. Parsa and D.G. Jamieson, "Identification of pathological voices using glottal noise measures" *J Speech Lang Hear Res* vol.43, pp. 469-485, 2000.

[8] J.R. Deller, J.G. Proakis and J.H.L. Hansen, *Discrete-time processing of speech signals*, 1993, Macmillan.

[9] N. Umeda, "Vowel duration in American English", *J Acoust. Soc. America*, vol.58, n.2, pp.434-445, 1975.

[10] N. Umeda, "Consonant duration in American English", *J Acoust. Soc. America*, vol.61, n.3, pp.846-858, 1977.

[11] J. Hillenbrand and R.A. Houde, "Acoustic Correlates of Breathy Vocal Quality", *J Speech Lang Hear Res* vol.39, pp.311-321, 1996.

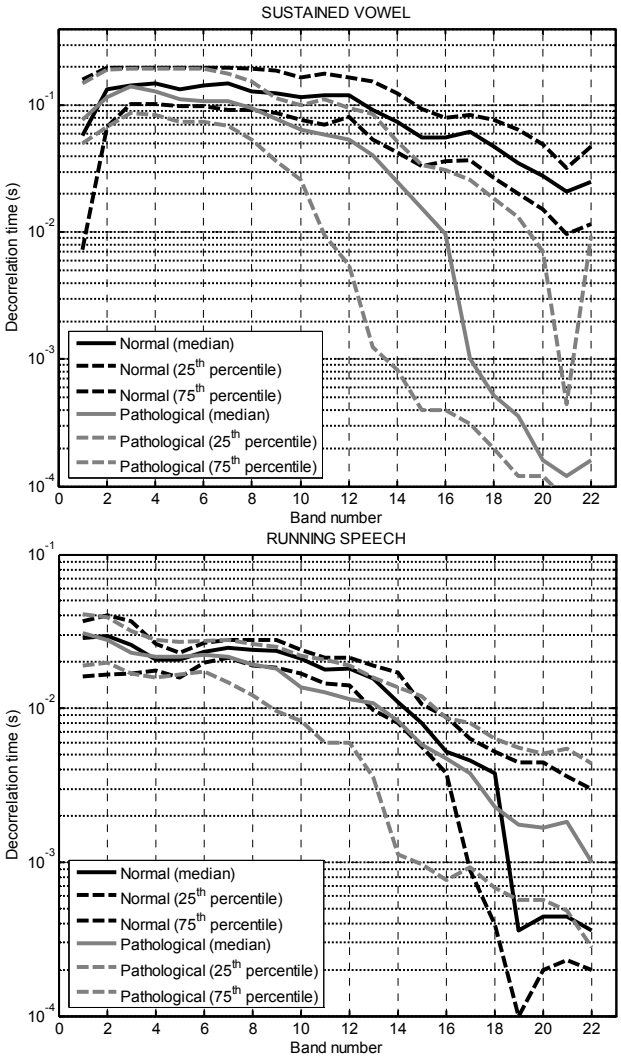


Figure 4. Distribution of decorrelation times of band energy sequences for sustained vowels (top) and running speech (bottom). Figure key is as in Fig. 3.

SENTENCE MODALITY RECOGNITION IN DYSARTHRIC SPEECH

D. Torres[1], T. Dekens[2], H. Martens[3], G. Van Nuffelen[3], M. De Bodt[3], W. Verhelst[2], C. A. Ferrer [1]

[1] Research Center on Electronics and Information Technologies (CEETI), Central University of Las Villas, Cuba
[2] Interdisciplinary Institute for Broadband Technology, Dept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Belgium
[3] Antwerp University Hospital, Rehabilitation Centre for Communication Disorders, Belgium

Abstract: **The ultimate goal of this research is to develop a tool for the automatic assessment and treatment of intonation and stress in dysarthric speech. In this paper, we deal with automatic sentence modality recognition in dysarthric speech. Two classes of sentence modalities were used: declarative statements and declarative questions. Statistics of prosodic features were used for the classification. Three well-known classification algorithms were tested with two different sets of features. The database used consisted of healthy and dysarthric speakers pronouncing three different sentences in both modalities. The healthy speakers were used as the training set and the dysathric speakers as the test set. A global classification accuracy of 84% has been achieved.**

Keywords: **dysarthric speech, intonation, pitch, energy**

I. INTRODUCTION

Dysarthria is a collective name given to a group of speech disorders caused by degenerative damage in the components of the peripheral or central nervous system. These disorders affect a large population, especially adults, and are commonly observed in neurological diseases [1].

Intonation is a prosodic element that gives information about the distinction among the different types of utterances such as imperatives, declaratives and exclamations. Intonation also conveys information about the speaker's state of mind. Intonation is related to changes in the fundamental frequency, intensity and timing of speech. The combination of different acoustic features gives information about the intonation modality [2].

Some research related with sentence modality recognition has been reported for different purposes such as automatic speech recognition [3] and prosodic assessment of language in hearing impaired children [4][5], among others. Previous research dealing with the characterization of intonation in dysarthric speech has been reported [6] but there are still no appropriate methods for the automatic assessment of intonation. Moreover, in human assessment, speech therapists often only indicate that the speech is monotonous or that the patient produces deviant intonation patterns.

The objective of this work is to explore the feasibility of using statistic measures of prosodic features to automatically assess the degree to which a patient can successfully produce an intended intonation pattern. Such information could be helpful for speech pathologists to make a diagnosis of the dysarthric patient's intonation efficiency. In this paper, we test a system for assessing whether a patient can produce the intonation for questions by building a classifier that can discriminate between a declarative question and the corresponding declarative statement in the Dutch language. The classifier is trained on speech from healthy speakers and the evaluation is done on speech from dysarthric speakers.

II. MATERIALS AND METHODS

A. Data

The recordings used in this work were acquired by the authors on healthy (Control Group, CG) and dysarthric adult patients (Target Group, TG) from the Antwerp University Hospital, Belgium (UZA). 13 healthy speakers and 20 dysarthric speakers were recorded. Three utterances were recorded for each type of sentence modality (declarative question, DQ and declarative statement, DS intonation). The recorded sentence pairs are:

1. Karen speelt tennis? / Karen speelt tennis. (Karen plays tennis? / Karen plays tennis.)
2. Hij kocht een jas? / Hij kocht een jas (He bought a jacket? / He bought a jacket.)
3. Je hebt de lotto gewonnen? / Je hebt de lotto gewonnen. (You won the lottery? / You won the lottery.)

Each sentence from both CG and TG was perceptually classified in DQ or DS by four experienced speech pathologists on dysarthric speech from UZA. Some recordings were excluded from the experiments due to the lack of inter rater agreement, yielding a final set of 76 sentences from normal speakers (41 DS and 35 DQ) and 112 sentences from dysarthric speakers (82 DS and 30 DQ).

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

B. System Description

The general algorithm used in this work is described as follows:

1. Extract the pitch and energy contours.
2. Preprocess the pitch and energy contours (interpolation, smoothing, extraction of the voiced sections and normalization by the mean to reduce inter speaker variability).
3. Extraction of the last 200 ms. Tilt parameters, energy and pitch statistical features extraction [7].
4. Extraction of statistical features from the energy and pitch contour in the whole utterance.
5. Classification (1: DS; 2: DQ).

The removal of the silences in the beginning and ending part of the recordings was performed so the next procedures were done using only the uttered interval (containing both voiced and unvoiced segments). The extraction of the pitch and energy contours was performed using PRAAT [8]. The autocorrelation method [11] was selected to estimate the fundamental frequency. Preprocessing steps include interpolation for the reconstruction of unvoiced segments and smoothing of the contours. Contour normalization was also included for reducing inter-speaker variability.

Table 1 Summary of the set of global statistical features extracted from the F0 and Energy contours.

Feature	Description
Max	Value of the maximum
PosMax	Position of the maximum
Min	Value of the minimum
PosMin	Position of the minimum
DifMaxMin	Absolute value of PosMax-PosMin
FRange	Range of frequency or energy Max-Min
Mean	Value of the mean
Std	Standard Deviation
Skw	Third statistical moment
Kurt	Fourth statistical moment
Q1	First quartile
Median	Second Quartile
Q3	Third Quartile
IQRange	Inter-Quartile Range
IQRange-Std	Absolute difference between IQRange-Slope
Slope	First coefficient of the linear regression.

Traditional statistical features like maximum, minimum, mean, quartiles, etc. were extracted from the whole utterance as well as features related to the rise and fall connection (RFC) model (maximum and minimum position). A summary of the set of statistical features used in the proposed system is shown in Table 1. As the ending voiced part of the intonation contour is indicative of the sentence modality, the last 200 ms of the utterance was extracted using PRAAT [8]. The intonation events in this segment (a: accents, b: boundaries) were described by means of three tilt parameters: initial Fo (Hz) at the start of the event, amplitude of the Fo excursion of the event (Hz) and tilt [9]. The Edinburgh Speech Tools Library [10] was the software used for the tilt parameters extraction. Statistical features related with the pitch and energy contours were also extracted from this segment, such as mean, maximum, and slope.

For the evaluation of the features' discriminative power and feature selection, five algorithms included in Weka [8] were used (BestFirst, Genetic Algorithm, Ranked Search, Linear Forward Selection and Random Search) [8] in a 10-fold cross validation. All the methods produced the same five features (4 features related to the pitch contour of the whole utterance and 1 related to the pitch contour of the last 200ms voiced segment of the utterance), as listed in Table 2. Features related with the energy contour were not selected by the five applied attributes selection methods.

Table 2 Features selected by the automatic algorithms

Feature	Description
Related with whole Fo contour	
Max	Value of the maximum
PosMax	Position of the maximum
PosMin	Position of the minimum
Slope	First coefficient of the linear regression.
Related with the Fo contour of the last 200ms	
Slope	First coefficient of the linear regression.

The classifiers tested for this application were:

1. Support Vector Machine (SVM) [13] using a polynomial (inhomogeneous) kernel represented by following expression:

$$K(x_i; x_j) = (x_i^T x_j + 1)^d \quad (1)$$

Where d is the exponent of the expression. Each x_i is a p-dimensional real vector that belongs to class y_j. For training the system Platt's Sequential Minimal Optimization (SMO) algorithm was used [14].

2. Decision Table (DT) using Best First as search algorithm [15].
3. Decision Tree J48 [16].

III. Results and Discussion

The algorithms mentioned above were applied on two sets of attributes: the full set of 41 features (16 statistical measures taken over both the whole pitch and energy contours, 3 statistical measures taken over both the pitch and energy contour of the last 200ms voiced part and 3 tilt parameters taken only over the pitch contour of the

last 200ms voiced part) and the set of five features obtained after feature selection. The training was done on the healthy speakers group and the classification test was performed on the dysarthric group.

A. Full set of features

Table 3 shows the confusion matrix for the SVM classification using the full set of 35 features.

Table 3 Confusion Matrix for SVM

Sentence Modality	DS	DQ
DS	67	15
DQ	10	20

The global accuracy of this experiment is 77% of correct classification. Only 25 instances out of 112 were misclassified. From these cases 10 declarative questions were misclassified as declarative statements and 15 declarative statements were misclassified as declarative questions. As we can observe in Table 3 with the use of the full set of features in SVM declarative statements are better recognized than declaratives questions.

Results for the experiments using DT are shown in Table 4. The numbers reveal that 94 cases out of 112 were correctly classified outperforming the previous results. From the 18 misclassified instances 5 declarative statements were misclassified as declarative questions and 13 declarative questions were misclassified as declarative statements.

Table 4 Confusion Matrix for DT

Sentence Modality	DS	DQ
DS	77	5
DQ	13	17

Table 5 Confusion Matrix for J48

Sentence Modality	DS	DQ
DS	75	7
DQ	15	15

The global accuracy of correctly classified instances using the J48 classifier was 80%. Table 5 shows the confusion matrix for this experiment. There were 22 misclassified instances and the class with the maximum number of errors was declarative questions with 15 errors (50% of DS). Globally, this classifier also outperforms the SVM classifier but it remains inferior to the DT classifier.

The result from experiments where the full set of features is used for classification thus reveals the superiority of the DT classifier with an accuracy of 83% correctly classified instances.

B. Reduced set of features after feature selection

As mentioned before attribute selection (AS) was performed using five well-known algorithms. The results reveal that only five features related with the fundamental frequency contour and the slope of the last voiced part of the utterance were the most relevant attributes for the classification of the sentence modality.

Table 6 shows the confusion matrix from the results of applying SVM classification. 95 instances out of 112 were well classified representing 84% of the total number of utterances. Only 4 declaratives questions were wrongly classified as declarative statements but 13 declaratives statements were mistaken for declaratives questions. In this experiment, in contrast to the one with the full set of features, the declaratives questions are better recognized than the declarative statements.

Table 6 Confusion Matrix for SVM-AS

Sentence Modality	DS	DQ
DS	69	13
DQ	4	26

Table 7 Confusion Matrix for DT-AS

Sentence Modality	DS	DQ
DS	72	10
DQ	13	17

At first sight, unexpected results were obtained in the experiments using the DT classifier. In Table 7, the numbers show that the amount of errors increases from 18 to 23 with respect to the experiments using the full set of features. In this case the number of correctly classified instances was only 89 (79%). One possible reason for this result could be that the DT classifier has its own feature selector for constructing the model. When the full set of attributes was used, from 306 subsets of features the optimal subset for classification using DT was obtained. The selected attributes used for DT-AS are different from those estimated by DT itself and therefore the set of attributes obtained from the feature selection algorithms is suboptimal for the DT classifier. Thus, the version which uses the full set of attributes shows better results than the version with the attribute selection.

The results from the application of J48 to the subset of selected attributes are to the same as the results of J48 using the full features set. In both cases the decision tree has the same structure. The features used by J48 in both cases are represented in Fig. 1. It can be observed how the J48 classifier only uses information about the slope of

the utterance and the slope of the last 200 ms to predict the sentence modality pattern.

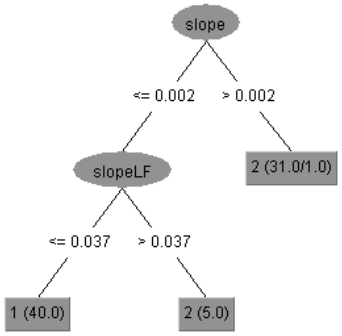


Fig. 1 J48 structure. It was obtained using both full features set and attribute selection subset

In the experiments related to the use of the selected attributes for the classification the best results were obtained by the SVM-AS method (84% of well classified instances).

V. Conclusion

This study addressed the design of an automatic system to help in the intonation assessment of dysarthric speech. Features related with the pitch and energy contours were extracted for characterizing the intonation modality. Three different classifiers were compared in this study using two different sets of features (full set and selected subset). In the use of the full set of features DT algorithm has the best results but the well-known SVM outperforms these results for the sentence modality detection in dysarthric speech using a reduced subset of attributes with a global accuracy of 84% of correct classification.

VI. Acknowledgement

The work reported on in this paper was performed while the first author was a visiting researcher at Vrije Universiteit Brussel with a scholarship from the Vlaamse Interuniversitaire Raad – VLIR (Flemish Interuniversity Counsel). This work is part of the project Computerized Assessment and Treatment of Rate, Intonation and Stress (CATRIS – IWT TBM-080662) that is supported by the Flemish government agency for Innovation by Science and Technology – IWT.

References

[1]. Darley, F.L., Aronson, A.E. & Brown, J.R. (1969). Differential diagnostic patterns of dysarthria, J. Speech Hear. Res., vol. 12, 246-249

[2]. Kohler, K. (2007). Beyond laboratory phonology – the phonetics of speech. In M. Sole, P. Beddor, & M. Ohala, Experimental approaches to phonology (pp. 41-53).Oxford: Oxford University Press. Pell, M., Cheang, H., & Leonard, C.

[3]. H. Wright (1998). Automatic utterance type detection using suprasegmental features, Proceedings of the International Conference on Spoken Language Processing

[4]. Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., Plaza, M. (2010). Automatic Intonation Recognition for the Prosodic Assessment of Language Impaired Children IEEE Transactions on Audio, Speech, and Language Processing, vol. PP, issue. 99, 1-16

[5]. Král, P., Klečková, J. & Cerisara, C. (2005). Sentence Modality Recognition in French based on Prosody World Academy of Science, Engineering and Technology, vol. 8, 185-188.

[6]. Joan K.Y. Ma , Rüdiger Hoffmann (2010). Acoustic Analysis of Intonation in Parkinson's Disease, Interspeech 2010, 2586-2589.

[7]. Paul A. Taylor. (1998). Analysis and synthesis of intonation using the tilt model. Proceedings of ICSLP88

[8]. Boersma, P. (2001). Praat, a system for doing phonetics by computer, *Glot International*, vol. 5, No. 9/10, 341-345.

[9]. Taylor, P. (1998) Analysis and synthesis of intonation using the tilt model. *Proceedings of ICSLP88*.

[10]. Edinburg Speech Tools, Available: http://www.cstr.ed.ac.uk/projects/speech_tools/

[11]. Boersma, P (1993) "Accurate short term analysis of the fundamental frequency and the harmonics to noise ratio of a sampled sound" IFA Proceedings, 17. 97-110.

[12]. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1

[13]. Hearst, M.A., Dumais, S.T. Osman, E. Platt, J. & Scholkopf, B. (1998). Support Vector Machines, Intelligent Systems and their Applications, IEEE, vol.13 issue.4, 18-28.

[14]. S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation. 13(3):637-649.

[15]. Ron Kohavi: The Power of Decision Tables. In: 8th European Conference on Machine Learning, 1995, 174-189.

[16]. Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

USING ACOUSTIC MEASURES TO PREDICT AUTOMATIC SPEECH RECOGNITION PERFORMANCE FOR DYSARTHRIC SPEAKERS

Kinfe T. Mengistu[1], Frank Rudzicz[1], Tiago H. Falk[2]
[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Institute National de la Recherche Scientifique, Montreal, Canada

Abstract: **There is growing evidence that clinicians are becoming more receptive to automated computerized tools that assist in treatment decisions and outcomes. Automatic speech recognition (ASR), for example, has had some degree of success as an assistive technology (AT) tool for individuals with mild or moderate dysarthria. Notwithstanding, for a large percentage of individuals with more severe levels of the disorder, ASR has yet to achieve acceptable levels. In this paper, we explore the use of several acoustic measures as correlates of ASR performance for dysarthric speakers. By automatically predicting the potential efficacy of ASR for a particular dysarthric speaker, health care costs and waiting lists may be reduced as may device abandonment rates. Experiments with the "Universal Access" database of dysarthric speech suggest that some of the proposed measures achieve correlations as high as 0.86 with ASR accuracy.**

I. INTRODUCTION

Speech is an efficient modality of communication in human-to-human interaction and can also serve as a high-capacity medium in human-machine interaction. However, millions of individuals have severe motor impairments that make speech communication extremely difficult, or even impossible [3]. These neuro-motor impairments, collectively known as dysarthria, are characterized by uncoordinated and imprecise articulation, and atypical breathing, voicing, and prosody that result in a highly distorted and unintelligible speech. Dysarthria is often accompanied by other physical handicaps that inhibit other forms of physical activity making the use of one's voice highly desirable. Recent advances in the automatic recognition of dysarthric speech have demonstrated that many individuals with speech disorders can be reasonably understood with specialized recognition software [4]. However, there remain many individuals with dysarthria for whom automatic speech recognition (ASR) remains insufficient, and for whom alternative forms of assistive technology (AT) need to be prescribed.

Being able to accurately predict the success of ASR by automatically analyzing a patient's speech signal would significantly expedite the AT prescription process, whilst also reducing device abandonment rates. This paper describes a number of acoustic measures which have been used in the past to objectively characterize the quality and intelligibility of both healthy [1] and dysarthric speech [2]. The goal is to explore the usefulness of each parameter as a correlate of ASR performance. It is known that dysarthria affects articulation, breathing, voicing, and prosody, often resulting in unintelligible speech. Therefore, we consider acoustic features that characterize the atypical vocal tract shape, vocal source excitation, temporal dynamics, and prosody characteristics of dysarthric speakers. More specifically, we explore the use of internal features computed by the speech quality measurement algorithm ITU-T Rec. P.563, standardized by the International Telecommunications Union (ITU) [1]. While the algorithm has not been optimized for dysarthric speech, some of its internal features may be useful for this task as they measure parameters related to atypical vocal tract shapes as well as atypical linear prediction coefficient (LPC) distributions.

Moreover, we explore the use of novel acoustic parameters proposed recently for the purpose of objective intelligibility prediction of spastic dysarthric speech [2]. These new parameters characterize atypical vocal source excitation, disordered temporal dynamics, and disrupted prosody, factors which are prominent in dysarthria. Here, we provide a brief description of the innovative features; the interested reader is referred to [2] for further details. Experiments with a publicly-available speech database show the acoustic measures investigated here as strong correlates of ASR performance on dysarthric speech. Such findings suggest that these measures can be used to predict the potential efficacy of ASR for disordered speakers, thus helping clinicians to better prescribe AT.

II. METHODS

A. Data Description

For computing the acoustic measures, we use a subset of dysarthric speech from the publicly available Universal Access (UA) Speech database from the University of Illinois at Urbana-Champaign. These data consist of single-word utterances recorded from 9 speakers (2 female) with spastic dysarthria recorded with a seven-channel microphone array, sampled at 16 kHz and digitized with 16-bit precision. Since the ITU-T P.563 standard requires single-channel narrowband (i.e., 8 kHz sampled) speech data, we further downsample the UA-database and use data from the sixth channel in the microphone array. This microphone was selected as it was placed closer to the participant and had a higher signal-to-noise ratio.

Each participant read 455 unique isolated words with some repetition totaling 765 utterances per participant. The prompts consisted of repetitions of English digits, the 26-word international radio alphabet, 19 word-processing commands, and the most common 100 words in the Brown corpus of written English. Each of these is repeat-

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

ed three times by each participant. In addition, 300 uncommon words selected from children's novels digitized by Project Gutenberg are also included [5].

B. Automatic Speech Recognition (ASR)

Baseline ASR performance is evaluated using speaker-independent (SI) acoustic models trained via the leave-one-out method where data from all speakers except the test speaker are used for training. The trained model is then evaluated on data from the test speaker. Each SI model is trained on an average of over 8000 dysarthric utterances. The acoustic feature vectors consist of 13 Mel-frequency cepstral coefficients (MFCCs) including the 0^{th}-order cepstral coefficient and their respective Δ and $\Delta\Delta$ coefficients, giving 39 dimensions generated every 10 ms. Cepstral mean subtraction (CMS) is then applied.

Acoustic models consist of 40 left-to-right, tri-state monophone hidden Markov models and a single-state short-pause model where state observation likelihoods are modeled by mixtures of 16 Gaussians. In each case, monophones are strung together into word networks according to the CMU pronunciation dictionary. A word-network where every word is preceded and followed by a silence model is used as a language model/task grammar. During decoding, a modified Viterbi algorithm is used to select the most probable word. All ASR accuracy results are reported in terms of word accuracy.

C. Acoustic Measures

A number of salient acoustic measures have been previously shown to characterize the quality of natural speech [1] and the intelligibility of dysarthric speech [2]. Below, a brief description of the measures are given; the interested reader is referred to [1,2] for more details.

C.1 ITU-T P.563 Algorithm

The ITU-T P.563 standard algorithm [1] was developed for narrowband telephone speech. As such, it detects and characterizes three major classes of telephone speech distortions, namely, background noise (both additive and multiplicative), temporal distortions (mute, clippings, interruptions) and unnaturalness (robotization and unnatural male and female speech). While the first two classes do not directly relate to dysarthric speech, we hypothesize that internal features computed by the algorithm and used to detect and characterize "unnatural speech distortions" may be useful for the task at hand. More specifically, the algorithm makes use of speech statistics for unnatural voice detection, such as higher-order statistical evaluation (kurtosis and skewness) of cepstral and linear prediction analyses. These are classical measures of the degree to which a statistical signal deviates from the Gaussian distribution. Kurtosis measures the 'peakedness' of a distribution and skewness measures the asymmetry of a distribution. Linear prediction analysis of order 21 is performed and kurtosis and skewness measures are computed for active speech.

We also consider five alternate acoustic parameters which were recently shown to correlate with subjective intelligibility ratings of spastic dysarthric speakers. The measures are based on three so-called intelligibility dimensions, namely atypical vocal source excitation, perturbation in speech temporal dynamics, and prosodic disruptions, as described below.

C.2. Vocal source excitation and vocal tract information

Linear prediction analysis has been widely used in speech applications to separate vocal source excitation and vocal tract information from the produced speech signal. Linear prediction analysis assumes that the current signal sample can be predicted by a linear combination of p previous samples. Under this format, the linear prediction error (or LP residual) will correspond to the vocal source excitation signal [6]. It is known that for healthy voiced speech segments, glottal pulses will appear as impulse-like peaks in the LP-residual signal, thus rendering the LP-residual distribution with a higher kurtosis [7]. On the other hand, severely dysarthric speech exhibits more prominent noise-like excitation signals (due to vocal harshness, for example), thus lowering the kurtosis value of the LP-residual distribution [2]. For mild to moderate dysarthric speech, it is expected that the kurtosis of the LP-residual distribution will lie between that of a Gaussian and that of healthy natural speech. For the sake of completeness, the LP-residual kurtosis metric κ is computed according to:

$$\kappa_{LP} = \frac{N \sum_{n=1}^{N} (r(n) - \bar{r})^4}{\left(\sum_{n=1}^{N} (r(n) - \bar{r})^2 \right)^2} - 3 \ ,$$

where $\bar{r}$ indicates the sample average of the LP-residual signal r(n) and N is the number of active speech frames.

C.3 Disturbances in temporal dynamics

Both short- and long-term temporal dynamics measures are explored to investigate the effects of temporal disturbances of spastic dysarthric speech on ASR performance. Speech temporal disturbances are mainly due to improper placement of the articulators, slower speech rate, and rhythmic disturbances [8]. Here, a log-energy rate of change measure is used to characterize the short-term temporal dynamics of the speech signal. More specifically, the $zero^{th}$-order cepstral coefficient c_0 is computed as a measure of short-term log-spectral energy and the $zero^{th}$-order delta coefficient Δc_0 is used as a measure of rate of change of log-energy [9]. In our simulations, c_0 is computed over 32 ms frames with 10 ms frame shifts and Δc_0 is computed using a window of size 7.

Statistics of the Δc_0 distribution are used to characterize disturbances in short-term (~100 ms) temporal dynamics. More specifically, the skewness computed from

C samples of Δc_0 distribution (represented by x_i in the equation below) is used:

$$S_{\Delta} = \frac{\sqrt{C}\sum_{i=1}^{C}(x_i - \bar{x})^3}{(\sum_{i=1}^{C}(x_i - \bar{x})^2)^{3/2}},$$

where $\bar{x}$ indicates the sample average of x_i.

Long-term temporal dynamics information, in turn, is characterized by the rate of change of long-term (between 512 and 1000 ms) speech temporal envelopes. Such representation is often termed "modulation spectrum" and characterizes slow energy fluctuations associated with the movement of the lips, the jaw, and other speech articulators. Most of the useful linguistic information is in modulation frequency components between 1 and 16 Hz, with spectral peaks around 4 Hz [11]. In [2] it was hypothesized that prolonged phonemes, slower speech rates, and impaired co-articulation would cause a shift of the modulation frequencies to below 4 Hz. With more intelligible speech, the modulation frequency would spread across higher modulation frequencies as observed with natural speech [12]. The ratio of modulation spectral energy at modulation frequencies less than 4 Hz to modulation frequencies greater than 4 Hz was used to measure long-term temporal dynamics [2]. This parameter, termed low-to-high modulation energy ratio (LHMR) in [2], takes into account temporal disturbances of irregular speech, namely prolonged phonemes, slower speech rates, and impaired co-articulation of dysarthric speech. In order to emulate psychoacoustic precepts, an auditory-inspired modulation spectral representation is used where a 23-channel gammatone filterbank was used to emulate the processing of the cochlea and an 8-channel modulation filterbank was used to aggregate modulation frequencies into eight bands [12]. A complete detail of the signal processing steps involved in the computation of the LHMR measure can be found in [2].

C.4 Disordered prosody

Prosodic disturbances are one of the distinguishing factors of dysarthria and we explore how these correlate with ASR performance. Here, the range and variance of the fundamental frequency (F0) [14] are used as acoustic parameters that characterize disordered prosody. Pitch estimates are computed using the robust adaptive pitch tracker algorithm [15].

III. EXPERIMENTAL RESULTS AND DISCUSSION

Table 1 shows the correlation coefficients attained between the investigated acoustic measures and ASR percentage accuracy over all speakers. As can be seen, the acoustic features that characterize atypical vocal source excitation and unnaturalness of speech are highly correlated with ASR performance on dysarthric speech. The LP-residual and LPC kurtosis, along with LPC skewness show strong positive correlations with ASR performance, with coefficients ρ ranging between 0.81 and 0.86. As expected, the LP-residual of relatively intelligible speech has a much higher kurtosis value (e.g., for M14) than severely impaired speech (e.g., for F03).

By contrast, the short- and long-term temporal perturbation measures, namely S_{Δ} and LHMR, show more modest correlations with ASR performance, achieving a coefficient of 0.62. Moreover, the range and variance of the fundamental frequency (F0), which are used to measure prosodic disturbances, are shown to be strongly negatively correlated with ASR performance. Dysarthric speech is commonly considered monotone and "robotic," thus it would be reasonable to expect lower pitch variability and range in more severe cases of dysarthria (and consequently, lower ASR accuracy). The negative correlations, however, suggest otherwise. While these findings may seem counterintuitive, they corroborate those reported in [14] where the nature of prosodic disturbances was shown to vary with the severity of dysarthria. In particular, monotonicity was reported for mild dysarthric speakers only and higher pitch variation/range was observed for speakers with severe disorders.

IV. CONCLUSIONS AND FUTURE WORK

This work has demonstrated that the investigated acoustic measures can be indicative of the performance achieved with traditional isolated-word recognition systems. In particular, acoustic measures related to atypical vocal source excitation and unnaturalness were highly correlated with ASR performance. As such, these measures can be used to assist clinicians in assessing the potential utility of ASR systems for particular dysarthric patients. For example, if LPC analysis of a patient's speech indicates LP coefficients with a high kurtosis, ASR systems are more likely to work as intended. In the future, a composite measure consisting of a weighted linear combination of these acoustic measures might further improve the predictive ability of this approach. Moreover, we are interested in further analysis of the relationships between specific motor disablements, spectral characteristics, and ASR performance. For example, prior research showed Pearson correlation coefficients of up to 0.95 between tongue motion and F2 formants for sonorants uttered by dysarthric speakers [4].

V. ACKNOWLEDGEMENTS

This research project was funded by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto. The authors thank Mark Hasegawa-Johnson for sharing the UA-Speech database.

Table 1. Correlation ρ between investigated acoustic measures and ASR accuracy for 9 dysarthric speakers.

Speaker ID	ASR % Accuracy	LP-Residual Kurtosis	LPC Kurtosis	LPC Skewness	S_Δ	LHMR	F0-Range	F0-Variance
F03	7.99	0.19	0.47	0.07	0.08	8.60	144.51	38.06
F05	34.80	1.22	4.32	0.48	0.44	5.04	121.23	36.37
M01	7.11	0.56	1.38	0.15	0.59	6.81	148.04	32.18
M04	3.39	0.36	1.21	0.18	0.31	6.86	122.96	30.56
M05	35.91	0.77	1.93	0.35	0.94	5.07	54.9	11.51
M07	21.41	0.38	1.30	0.14	0.44	9.20	116.79	28.96
M08	61.94	0.98	4.51	0.81	0.93	5.91	73.97	18.81
M14	50.49	1.29	5.49	0.93	0.57	4.80	27.29	6.43
M16	33.39	0.80	1.73	0.14	0.23	6.46	129.55	30.34
ρ coefficient		**0.81**	**0.84**	**0.86**	**0.62**	**-0.62**	**-0.76**	**-0.67**

REFERENCES

[1] ITU-T P.563, 2004. *Single-ended method for objective speech quality assessment in narrow-band telephony applications*.

[2] Falk, T., Chan, W.-Y., Shein, F., 2011. Characterization of atypical vocal source excitation, temporal dynamics and prosody for machine measurement of dysarthric speech intelligibility. *Speech Communication*, in press.

[3] Selouani, S-A., Yakoub, M.S., O'Shaughnessy, D., 2009. Alternative speech communication system for persons with severe speech disorders. *EURASIP Journal on Advances in Signal Processing*, Vol. 2009.

[4] Rudzicz, F., 2011. *Production knowledge in the recognition of dysarthric speech.* PhD thesis, University of Toronto.

[5] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. *Proceedings of INTERSPEECH 2008.*

[6] Deller Jr., J. R., Proakis, J. G., and Hansen, J. H. L., *1993. Discrete-Time Processing of Speech Signals*. New York: MacMillan.

[7] Gillespie, B., Malvar, H., Florencio, D., 2001. Speech dereverberation via maximum-kurtosis sub-band adaptive filtering. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* pages 3701—3704.

[8] Duffy, J. R., 2005. *Motor speech disorders. Substrates, differential diagnosis, and management* (2nd ed.) . Mosby, St. Louis.

[9] Huang, X., Acero, A., Hon, H.-W., 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice-Hall, New Jersey.

[10] Picone, J., 1993. Signal modeling techniques in speech recognition. *Proceedings of the IEEE* Vol. 81, No. 9, pages 1215—1247.

[11] N. Kanedera, N., Arai, T., Hermansky, H., Pavel, M., 1997. On the importance of various modulation frequencies for speech recognition. *Proceedings of Eurospeech '97*, pages 1079—1082.

[12] Drullman, R., Festen, J. , Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America,* Vol. 95, No. 5, pages 2670—2680.

[13] Dau, T., Puschel, D., Kohlrausch, A., 1996. A quantitative model of the effective signal processing in the auditory system. I – model structure. *Journal of the Acoustical Society of America*, Vol. 99, No. 6, pages 3615—3622.

[14] Schlenck, K., Bettrich, R., Willmes, K., 1993. Aspects of disturbed prosody in dysarthria. *Clinical Linguistics & Phonetics*, Vol. 7, No. 2, pages 119—128.

[15] Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT), in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pages 495—518, Elsevier Science Publishers, Amsterdam.

VOICE QUALITY ANALYSIS TO DETECT NEUROLOGICAL DISEASES

[1]P. Gómez, [1]V. Rodellar, [1]V. Nieto, [1]L. M. Mazaira, [1]C. Muñoz, [2]M. Fernández, [2]E. Toribio
Grupo de Informática Aplicada al Procesado de Señal e Imagen, Universidad Politécnica de Madrid
[1]Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
Tel.: +34913367384, fax: +34913366601, e-mail: pedro@pino.datsi.fi.upm.es
[2]ENT and Neurology Services, Hospital del Henares, Avda. Marie Curie s/n, 28822 Coslada, Madrid, Spain

Abstract: **Neurological degenerative diseases are becoming a growing concern in modern society. The successful treatment of these diseases depend greatly in early detection. Speech has been routinely used by specialists as a valuable correlate in the assessment of pathological disease. Specifically voicing can serve as a very introspective correlate for this practice. The present paper uses a methodology previously employed in organic pathology voice quality assessment to explore to what extent specific low-level correlates of neurological diseases may be established. The methodology uses voiced recordings of sustained vowels to estimate vocal fold visco-elastic parameters from inverse filtering. These parameters show to be clearly influenced by unstable neuronal spiking resulting in tremor which affects many phonation cycles. The possible modeling of tremor could be used as an index to neuro-motor problems in phonation and help in differential diagnose of the pathology at an early stage. The paper presents examples on parameter estimations from study cases of spasmodic dysphonia and Parkinson Disease. Further development of research lines on this estimation methodology is also addressed.**

Keywords: **Inverse Filtering, Vocal Fold Biomechanics, Parkinson Disease, Voice Quality Assessment, Tremor**

I. INTRODUCTION

Classically Voice Quality Analysis has been focused to detect and establish the organic pathology in voice resulting from pathological alterations of larynx physiology. The study of other sources of dysphonic voice finding their ultimate reasons in the alterations of the neurological paths controlling phonation have been tagged as "functional" or "non-organic". Voice resulting from altered phonation due to neurological reasons may be a most valuable report of the etiology and progress of neural diseases affecting the production of voice, such as pathologies resulting in voice tremor [1]. These would include some kinds of spasmodic dysphonia, stammering and Parkinson. The possibility of early detection in the first stages of Parkinson's Disease (PD) may grant a better preventive treatment reducing the progress of the illness [2]. Monitoring treatment by objective methods is also an important goal, especially in modifying or defining new protocols. The deepest foundations of the methodology proposed in this paper are to be found in tracking the malfunctioning of neurological and neuromuscular paths involved in voice production (see Fig. 1).

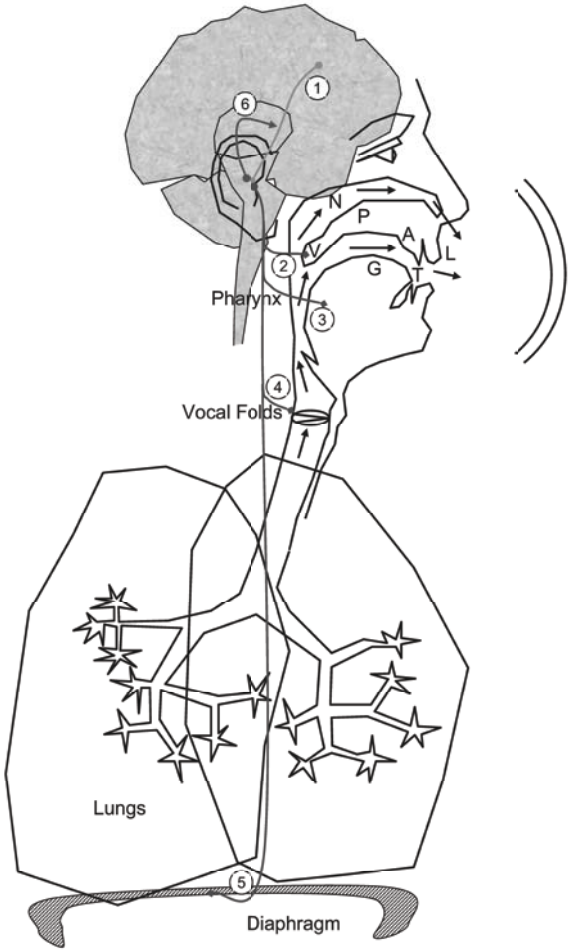


Fig. 1. Simplified view of main neural pathways involved in the production of phonation: 1. Links from linguistic neuromotor cortex to Basal Ganglion relay stages. 2. Branch of the X nerve acting on the naso-pharingeal switch. 3. Idem acting on the retro-lingual switch connected to the epiglottal switch. 4. Branch of the laryngeal nerve acting on the transversal and oblique arytenoid and cricothyroid muscles responsible for the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

vocal fold adduction and abduction. 5. Branch of the vagus nerve (phrenic) actuating on the diaphragmatic muscles. 6. Feedback loop in Basal Ganglia damping muscular tone.

These comprise links from the neuromotor linguistic cortex [3] to the subthalamic region [4] and through the laryngeal nerve and their associated pathways [5][6] to the muscles activating the thyro-arytenoid structure, responsible in the last term of vocal fold stretching, adduction and abduction (Superior Laryngeal Nerve, Internal and External Laryngeal Branches of the Inferior Laryngeal Nerve, Transverse and Oblique Arytenoid Muscles -TOAM-, and Cricothyroid Muscles -CM). Any alteration in the functionality of these pathways and in the associated muscles will result in temporary distortions of the parameters of tension and dynamic mass contribution of the vocal folds, both on the body and the cover biomechanics. Correlates of these alterations will be found in the pitch, and in long term jitter and shimmer, as the periodicity of these alterations may be of hundreds of milliseconds [7]. The aim of this paper is to give some phenomenological account in detecting and grading the neurological disease using biomechanical correlates obtained from the inverse filtering of voice. The technology has been tested in monitoring pre-post treatment of organic pathology, and due to its ubiquitous character can be applied as well to the neurological disease.

II. METHODOS

A database of voice recordings from neurological disease-affected patients is being recorded in Hospital del Henares of Madrid. This geographical area South East of Madrid is specially sensitive to PD. Being a heavy industrial area it is believed that some environmental factors may be responsible of the largest incidence of PD among the aging population compared to other regions of Madrid. For the preliminary and explorative character of the present study some specific cases are selected, these being strong spasmodic and PD voice samples, pathological voice of organic origin and voice from normophonic patients to serve as a contrast (all of them females). These voices are inverse filtered and some biometrical and biomechanical parameters are estimated, as the glottal closure sharpness, the mucosal/average ratio, the first two cepstral coefficients of the glottal source power spectral density, and the tension of the vocal fold body. It may be shown that these indices show a strong correlation with the spasmodic episodes both in their timely evolution and statistical dispersion. The methodology is based on the following steps:

1. Three emissions of the vowel /a/ are recorded at 44,100 Hz under normal phonation conditions.
2. For specific statistical comparison they are low-pass filtered and re-sampled to 22,050 Hz. High-pass filtering at 25Hz is also applied to eliminate low frequency flickering effects. Frames of 0.4 s long are used in the analysis.
3. Inverse Filtering is applied, and the glottal source is reconstructed [8].

Estimations of the glottal closure sharpness, noise/glottal ratio, dynamic mass and tension of the vocal fold body are derived following [8].

III. RESULTS

An episode of spasmodic dysphonia (SD) has been selected from the database to show the possibilities of the methodology, corresponding to a female voice (32 year old) manifesting about 2-3 spasms per second. The record is a segment of 0.4 s long from a sustained phonation of vowel /a/ (see Fig. 2 and Fig. 3).

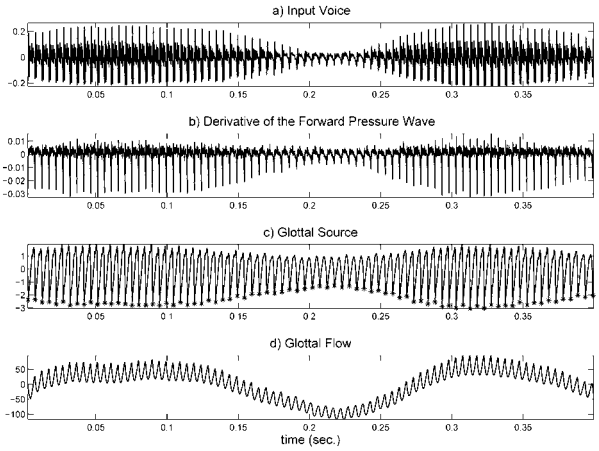


Fig. 2. Episode of spasmodic dysphonia. Templates from top to bottom: Voice signal. Inverse filtering residual. Glottal source. Glottal flow.

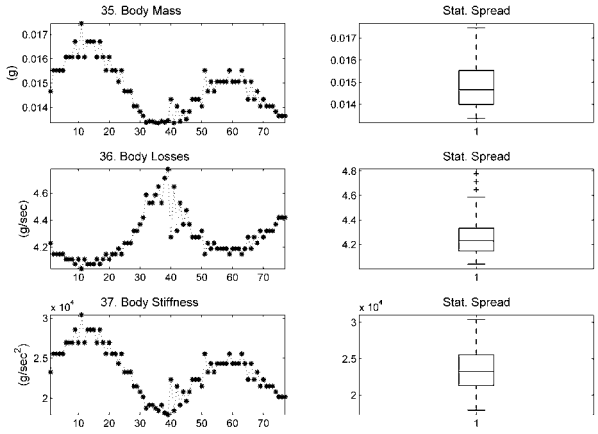


Fig. 3. Left templates from top to bottom: Phonation cycle-synchronous estimates of the dynamic mass component of the vocal folds, friction losses and body stiffness. Right templates from top to bottom: Statistical distributions of the left templates given as boxplots.

The voice segment studied is a part of a recording of a sustained /a/ 0.4 s long where an episode of spasm is clearly recognizable by the amplitude decay. The reconstruction of the glottal source (template c) does not show such a strong decay in amplitude. Pitch ranges from 208-203 Hz in the sections out of the spasm to a minimum of 185 Hz during the spasm, following an almost regular fluctuation (tremor of about 2.5 Hz). It may be seen that the estimates of the dynamic mass of the vocal fold body, and especially the fold tension are highly correlated with the spasm, reporting changes of about 25% and 50% of variation respectively. Similar fluctuations are found in other distortion parameters, such as the sharpness of the closure spike in the glottal source, the noise/glottal energy ratio and some cepstral parameters of the glottal source spectral density.

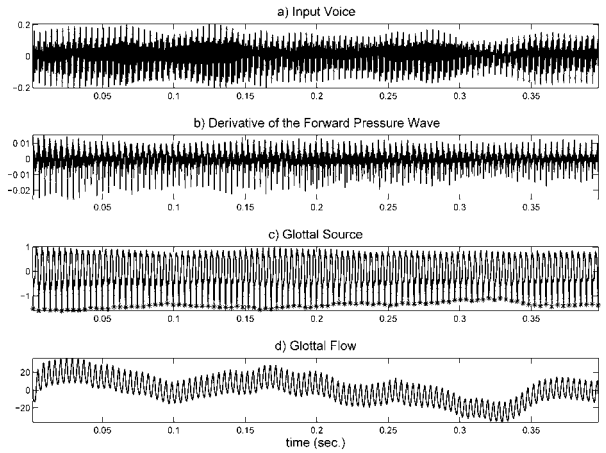


Fig. 4. Phonation 0.4 s long from a patient affected from Parkinson Disease. Templates from top to bottom: Voice signal. Inverse filtering residual. Glottal source. Glottal flow.

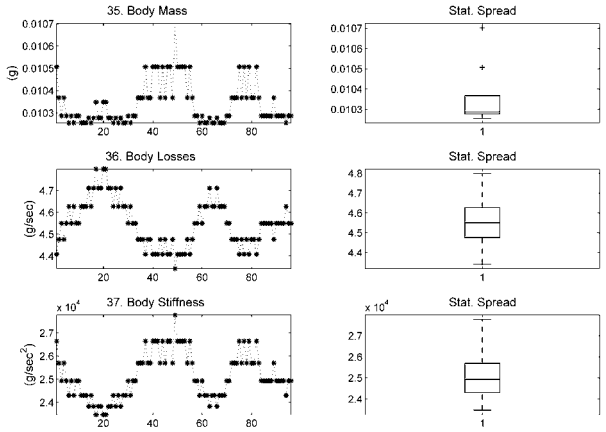


Fig. 5. Left templates from top to bottom: Phonation cycle-synchronous estimates of the dynamic mass of the vocal folds, friction losses and body stiffness. Right templates from top to bottom: Statistical distributions of the left templates given as boxplots.

A second example from a patient (72 year old) affected by Parkinson Disease (PD) corresponding also to female voice has been analyzed following the same methodology. The record is a segment of 0.4 s long from a sustained phonation of vowel /a/. The results of the analysis are reported in Fig. 4 and Fig. 5. In this case the changes in amplitude are not as relevant as in the spasmodic case. The reconstruction of the glottal source (c) does not show important changes in amplitude as well. Pitch ranges from 240-256 Hz following an irregular fluctuation (tremor) of about 5 Hz. The estimates of the vocal fold body dynamic mass and stiffness report changes of around 20% . To put the analysis into context at this point it would be worth to compare some overall results for these two cases against results from a normal female speaker and a pathological female speaker. The normal speaker (NF) is a 34 year old female, non-smoker not having reported any problem with voice, volunteering for the study. Normal condition was assessed by endoscopy and EGG. The case with organic pathology corresponded to a female 22 year old having been diagnosed from a left vocal fold cyst (LVFC) affecting the contralateral fold (contact lesion). The case was graded 2 (severe) in GRBAS scale. Endoscopy and EGG availed the diagnose. The acoustic processing of the four cases included the extraction of pitch, relative jitter and shimmer and the noise/glottal energy ratio (NGE). The stiffness of the vocal fold body was estimated as well. The results are given in Table 1 at the end of the paper.

IV. Discussion

From the results in Table 1 the first consequence is that pathological data (except for PD) are clearly differentiated from normal data in the value of the dispersion (standard deviation) and in the stiffness of the vocal fold body. Mean values of the classical distortion estimates as jitter, shimmer or NGE do not show important differences among the pathological cases except in PD. This case shows distortion parameters which could be considered normal. The problem is that tremor in PD is observed as FM-modulations which do not leave clues in the jitter, whereas the SD case may be traced in shimmer. Going to the causes, it seems that the effects of FM-affected spiking producing tremor in SD may be observed on the specific muscles affecting vocal fold abduction and adduction (TOAM-CM) as well as in the muscles responsible for pressure build-up and sustenance in lungs during phonation (diaphragm). The influence of FM-affected spiking in the modulation of the vocal tract (naso-velar switch, glossomuscular and oro-labial complexes) could also introduce changes in the production of voice, interfering with vocal-fold induced tremor. These differences may affect the results observed, as in the two cases studied. In the spasmodic case the

important changes in amplitude observed could be associated with some influence of the spasm on the diaphragm and other muscles inducing subglottal pressure, besides affecting strongly to the vocal fold stiffness as a result of the TOAM-CM action. The result during the spasm is a dystonic relaxation of the vocal folds (abduction) accompanied by a decay in subglottal pressure. The case of the PD patient may have to see only with the action of the TOAM-CM, resulting in a relatively cyclic dystonic behavior of the vocal fold but not in important changes of the subglottal pressure. It seems that parameters tracking amplitude changes as shimmer or APPQ measured directly on the glottal source, as well as the indirect estimates of vocal fold tension may serve as important marks to produce differential diagnose in tremor-affected dysphonias, and this line should be further studied. Other possible correlates could be the sharpness of the closure instant and the lowest cepstral coefficientes of the glottal source spectral profile. This means that the study of tremor as a result of neurodegenerative diseases may require complex time-frequency analytical techniques. Chaotic modeling of tremor in stiffness and other correlates, and Wavelet Transform may be good candidates out these studies.

V. Conclusions

The first conclusion from this phenomenological description is that tremor appears as a mark in certain biomechanical estimates of vocal fold dynamics as body stiffness. Therefore the monitoring and modeling of tremor could be based on the study of these correlates. Indications that differential diagnose could also be based in combined amplitude-stiffness indices are plausible enough for the issue to deserve further study. The analysis of the mentioned correlates estimated directly from the glottal source obtained after vocal tract inversion instead of whole voice may be a beneficial methodology to unveil and quantize the extent or degree of the spasmodic or tremor illness. As the characterization of tremor in voice shows quasi-cyclic information, techniques to model this characteristic as chaotic attractors, wavelets, or ARMA coefficients may be of much higher resolution than the analysis of full voice. The monitoring of neurological diseases is of most importance in a world where the aging of general population will demand important resources for health care. The early detection and monitoring of these problems may help in devising more efficient treatment protocols. Routine voice tests may help in this task. The validation of this methodology for PD is in due course in cooperation with the ENT and Neurology Services at Hospital del Henares.

References

[1] Pantazis, Y., Koutsogiannaki, M., Stylianou, Y., "A novel method for the extraction of tremor", *Proc. of MAVEBA07*, Florence University Press 2007, pp. 107-110.

[2] Das, R. “A comparison of multiple classification methods for diagnosis of Parkinson disease”, *Expert Systems with Applications*, Vol. 37, 2010, 1568-1572.

[3] Rauschecker, J. P., Scott, S. K., "Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing", *Nature Neuroscience*, Vol 12, 2009, pp. 718-724.

[4] Törnqvist, A. L., Schalén, L., Rehncrona, S., "Effects of different electrical parameter settings on the intelligibility of speech in patients with Parkinson's Disease treated with subthalamic deep brain stimulation", *Mov. Disord.*, Vol. 20, 2004, pp. 416-423.

[5] Eckley, C. A., Sataloff, R. T., Hawkshaw, M., Spiegel, J. R., Mandel, S., "Voice range in superior laryngeal nerve paresis and paralysis", *J. Voice*, Vol. 12, 1998, pp. 340-348.

[6] Luschei, E. S., Ramig, L. O., Baker, K. L., Smith, M. E., "Discharge characteristics of laryngeal single motor units during phonation in young and older adults and in persons with Parkinson disease", *J. Neurophysiol.*, Vol. 81, 1999, pp. 2131-2139.

[7] Tsanas, A., Little, M. A., McSharry, P. E., Ramig, L. O., "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests", *IEEE Trans. Biomed. Eng.*, Vol. 57, 2009, pp. 884-893.

[8] Gómez, P., Fernández, R., Rodellar, V., Nieto, V., Álvarez, A., Mazaira, L. M., Martínez, R, Godino, J. I., “Glottal Source Biometrical Signature for Voice Pathology Detection”, *Speech Communication*, Vol. 51, 2009, pp. 759-781.

Table 1. Parameter values for the four cases studied (standard deviations between parentheses)

Subject/Parameter	Pitch Hz	Jitter %	Shim %	NGE %	Stiffness ($g.s^{-2}$)
#346 (34y NF)	199 (1.15)	0.7 (0.6)	1.9 (1.3)	8.4 (0.5)	19542 (185)
#341 (22y LVFC)	215 (7.04)	4.2 (3.6)	3.8 (2.4)	6.6 (1.9)	24857 (4487)
#308 (45y SD)	199 (6.02)	1.5 (1.5)	6.5 (3.3)	8.9 (1.7)	22168 (2656)
#337523 (72y PD)	248 (3.87)	0.7 (0.5)	1.4 (1.0)	6.5 (1.1)	25138 (988)

Singing voice

Invited Speaker and introduction:

F. Fussi

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

THE VOCAL SCORE PROFILE/VOICE RANGE PROFILE RATIO (P/P RATIO) IN ARTISTIC VOICE EVALUATION: APPLICATION TESTED ON OPERA AND MUSICAL SINGERS

Franco Fussi[1], Nico Paolo Paolillo[2]

[1]Centro Foniatrico USL Ravenna, Teatro Comunale di Bologna, Ravenna, Italy
[2]ENT department Mandic Hospital (Merate-LC), Teatro alla Scala, Milan, Italy

Abstract: **Performances of unsuited repertories to singer's vocal and technical features can cause increasing risks of vocal effort (VE) and fatigue or glottis injury (GI), then it's important to find the right repertory for artist's vocal and technical features. We made manual voice range profiles (VRP) and interviews regarding performed, studied or not studied roles in professional singers. The dynamic agility (DA) curve, that is the differentials' curve (note by note) between loud and soft phonation curves of phonetogram, was obtained from VRP. This type of curve allows us to assess the phonation system capacity all range long. We realised for each operatic and musical role a vocal score profile (VSP), that is a statistic method for vocal score semeiotic and accurately highlights the vocal role various musical features through histograms and numeric parameters (1). Then we superimposed the DA graphs on VSP graphs creating a new graph (P/P ratio) that gives a synoptic summary of suitableness of examined singers' vocal and technical features in regard to considered role, revealing hard and critical moments eventually causing higher VE and GI risks (2). At last we compared data from P/P ratio with those from interviews, valuating correspondence between subjective and objective data (3).**
This study describes explicative examples of graphs analysis; in all cases analyzed through P/P ratio we found easiness in data interpretation, reliability in suitableness evaluation and expectation, good correspondence between subjective and objective data.

Keywords : **vocal score profile, voice range profile, partiturogram, phonetogram, vocal effort.**

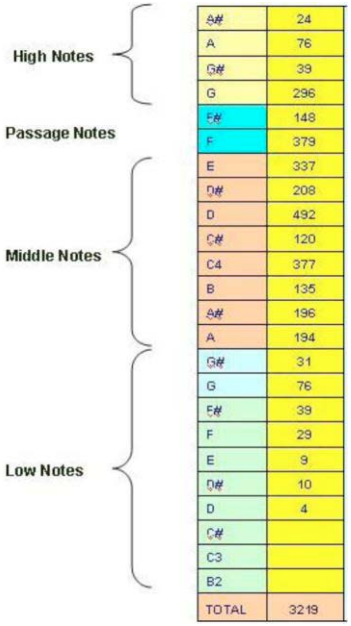
I. INTRODUCTION

In the clinical management of the artistic voice is important to identify all the risk components of vocal fatigue or glottic damage.
The choice of repertoires unsuited to technical and vocal features is one of factors that increase the risk of vocal effort and fatigue. For this reason it is essential to predict and to assess the vocal cost in performing a specific role to avoid any risk of glottic damage. We examined opera and musical singers using a method we have developed The first part of the investigation was to evaluate singers' feeling about characters' features making interviews before analysis to assess accordance between subjective feeling and objective analysis, interviews after analysis to evaluate the reliability of a predictive evaluation and an accurate anamnesis to find relationships between the patient's medical history, such as any vocal desease or phonosurgery, and results of the analysis.

II. METHODS

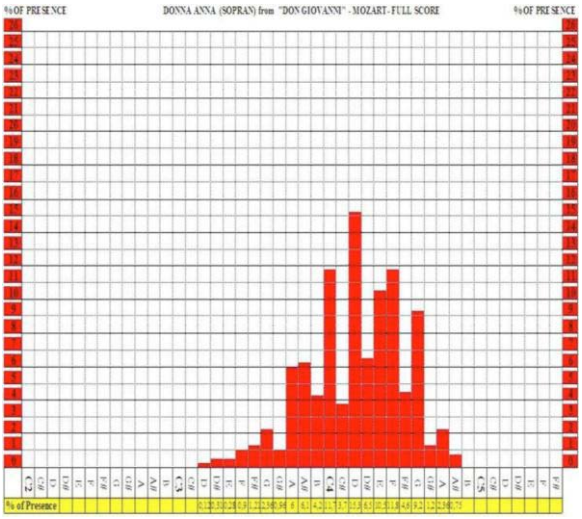
We used VOCAL SCORE PROFILE, a statistic method for semeiotic of complete vocal score or partitura. It's made counting presence of notes for each semitone in vocal score using the following scheme (on the top the duration of notes, on the left the tempo).
Here is an example regarding the role of Mozart's Don Giovanni: donna Anna.

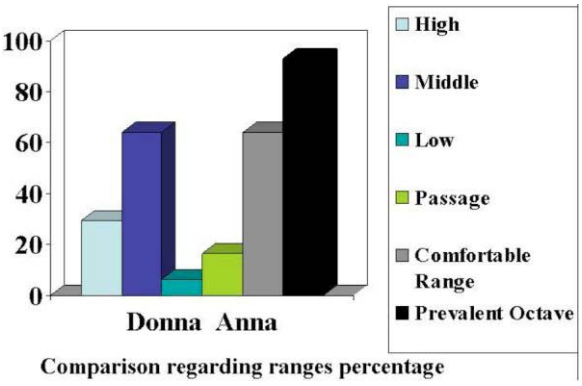
	𝅝 Semibreve	𝅗𝅥/2 Minim	♩x2 Crotchet	♩ Quaver	♪/2 Semiquaver	♪/3 Demisemi quaver	♪/4 Hemidemi semiquaver
Presto-Prestissimo (144-208) Vivace (126-144) Allegretto-Allegro (100-126)	2	1	1/2	1/4	1/8	1/16	1/32
Moderato (80-100) Andante-Andantino (60-80)	4	2	1	1/2	1/4	1/8	1/16
Lento-Adagio (50-60) Largo-Larghetto (44-50) Grave (40-44)	8	4	2	1	1/2	1/4	1/8



A#	24
A	76
G#	39
G	296
F#	148
F	379
E	337
D#	208
D	492
C#	120
C4	377
B	135
A#	196
A	194
G#	31
G	76
F#	39
F	29
E	9
D#	10
D	4
C#	
C3	
B2	
TOTAL	3219

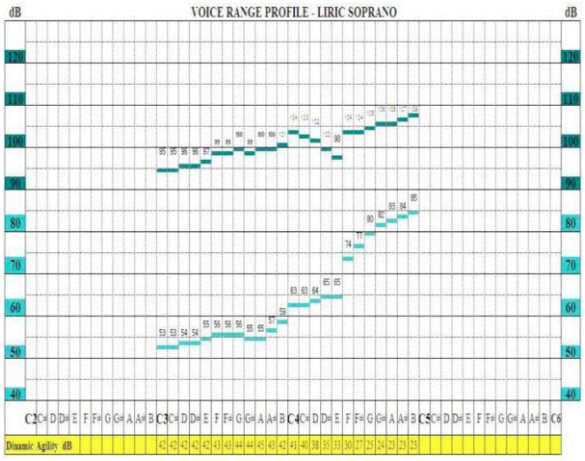
It is possible to obtain percentages of presences for every tonal range such as low, middle, passage, high, prevalent octave and comfortable ranges. Here are two istograms for the comparison of the same previous roles.
Using percentage of presence of notes for each semiton, we created an histogram called vocal score profile or partiturogram: below we can observe the tonal range semiton by semiton and on the right or on the left the percentage of presence.
Here we have vocal score profile for donna Anna role. In this case we can note an high presence of middle, passage and high notes , a prevalent octave from A3 to A4 and a little percentage of presence of comfortable range notes. So we can confirm that this is a really difficult role, suitable to a lyric soprano.



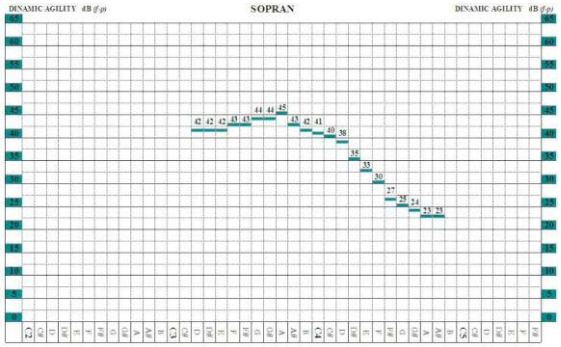


Comparison regarding ranges percentage

We made a voice range profile in a liric soprano, usual performer of this two roles in many important theatres under famous conductors. As you know VRP points out dynamic and frequency range of singers voice.

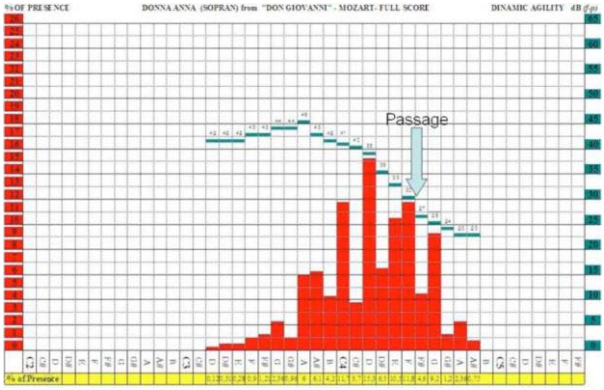


Anyway it's better to consider the Dynamic Agility, which is the value of differential between forte and piano calculated for each tone and allows to accurately value the phonation system capacity all range long. In this case we see a decreasing dynamic agility since middle tonal sector and a worsening in passage and high sectors



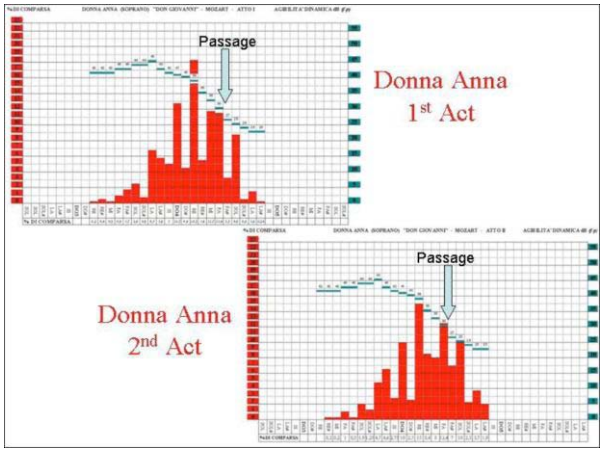
Superimposing the dynamic agility graphs on vocal score profile graphs, we obtain a graph (we called P/P ratio) that gives a synoptic summary of suitableness of examined singers' vocal and technical features in regard to considered roles, revealing the hardest and critical moments for the singer eventually causing higher vocal effort or injury risks. Below there is the tonal range, on the right the dynamic agility in dB and on the left the percentage of presence of notes.

III. RESULTS

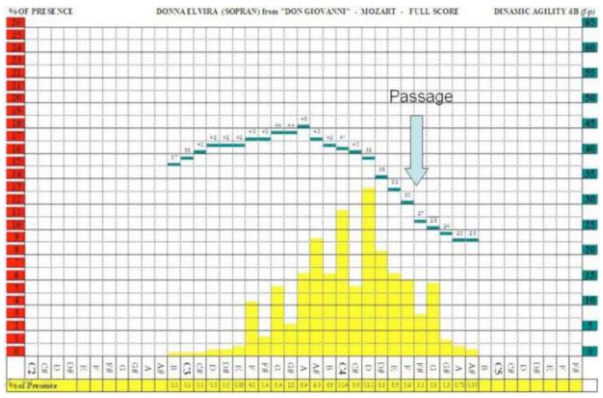


We note that in role critical zones, middle, passage and high ranges there's an evident reduction of dynamic agility, revealing an high risk of vocal effort. The singers noticed that it's a difficult role, expecially for high and passage notes and is more fatiguing than donna Elvira, the other soprano role in don Giovanni of Mozart. In singer's feeling, the first act is easier than the second.

If we analyze graphs for each act we can point out this feeling: in fact in second act there's an higher presence of passage and high notes, which make the part more difficult than in first act.



Here we have graphs regarding donna Elvira role: we clearly point out that the lower presence of passage and high notes makes the role more accessible and easier for the vocal features of this soprano.



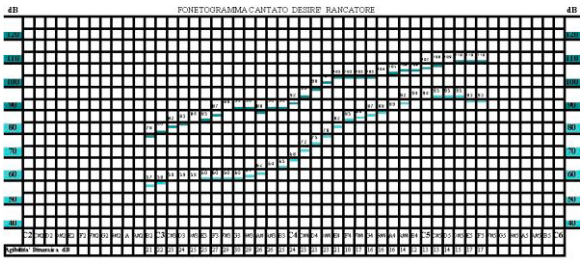
The lyric soprano previously analyzed has sung donna Anna role more than 1 hundred times through 3 years on stage in many important theatres under prestigious conductions, sign of a good performance in that role, but with the final result consisting in phonosurgery and a current vocal folds damage.

About donna Anna and donna Elvira roles in Mozart's Don Giovanni, she told us: "Donna Anna requests a certain vocal tract from the first to the second act. The first act is for a lyric soprano with a very dramatic temperament, a quite hard script, always touching a medium-high *tessitura* and consequently quite tiring. On the other hand the second act is completely different, everything becomes lighter, the *tessitura* becomes higher and requests a lighter vocal tract. The difficulty is just in finding the right balance, both vocal and physical, between the first and the second act: it is necessary not to give too much during the first act and equilibrate the second act with respect to the first one.

Referring to me, I felt more at my ease during the first act, perhaps also because I have a more full-blooded temperament. In the second act we can say that the thought of the second aria was warring me a bit; but obviously it is a marvellous part. I played it but at the end I felt a bit tired as if I had being using a bit too much the material and not the interest. On the contrary, during the first act I could even begin without vocalizing because I felt it as being mine. Further on I also sang the role of Elvira and I must say that I really felt at ease because it is a more natural script; it is more similar to the speaking way of a woman while Donna Anna represents the one who extremes the voice. In Donna Elvira both the recitative and ensemble parts are more natural and this because it is always written in a very natural tessitura. On the other hand Donna Anna is continuously in a medium / high-notes voice section also in the ensemble. This meaning that she must always sing in a low voice, in a very low voice. As a consequence who has not extreme facility in that zone can feel some tiredness, with respect to Donna Elvira who constantly remains coherent from the beginning to the end. It is role that must be sung with temperament, with expressiveness but for what was related to the vocal effort, to me the *weaving* resulted more comfortable".

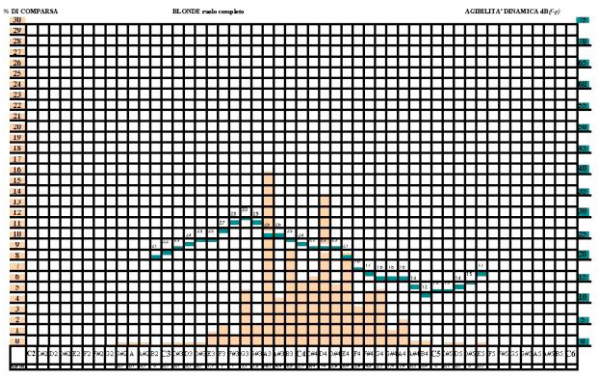
Finally we can see here the Phonetogram of the actually most famous coloratura soprano in Italy, Desirée Rancatore, known interpreter of Zauberflute, Die Entfuhrung aus der Serrail, Lakmè, Lucia di Lammermoor, Rigoletto and wonderful Doll in Les Contes d'Hoffmann.

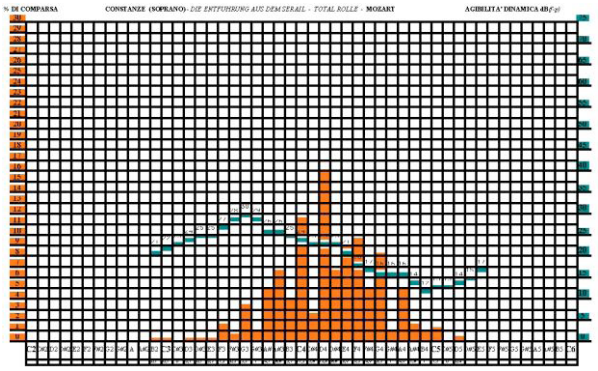
About Blonde and Constance roles in Die Entfuhrung she told us: "The difficulty of Blonde is in the *weaving*, as it is all central, in the tuned-up, especially those quite lower than Constance.



On the other hand, in the aria she is a pure colour-full soprano, with sudden natural high E. Therefore it is necessary a Pure colour-full soprano to get facility with high notes; while the difficulty is in facing low notes. In Constance the difficulty consists in standing the whole role, because it is long and technically difficult. In my opinion she is not at all a colour-full soprano just to play the role; it is a very dramatic and expressive part. Besides there are two consecutive arias, such as *Traurigkeit ward mir zum Loose* and *Martern aller Arten*, which are completely opposite, for one is central,

lyric and sad, while the other in impetuous and dramatically colour-full. In such a case it is difficult to stand the two arias length".
In the utmost Constanze aria, our soprano has major effort on the high tones C5 and Cdiesis5, as she shows lesser differential from the H4 to the D5, and –as a paradox- she shows more differential comfort on the high notes after D5.
The prevalent octave of the role is rather unbalanced towards high notes, in one range (A2-A4) that, according the phonetogram, is more suitable for the singer, both in the aria and even more in the whole role. Besides the singer is at ease in the passage notes and in the well represented centers in the part.





IV. CONCLUSION

As a conclusion the relationship between phonetography of a singer and partiturogram of one role, allows us to make some consideration, useful for the singer, about his adequateness to the role iself in function of the larynx muscular effort, phonastenic probabilities, and the forecast of major or lesser rest necessities among the various performances.
We conducted several tests of this type on many singers, 10 opera and 7 musical and we revealed that
- P/P ratio is a reliable method to identify a suitable repertoire and to predict performance risks for vocal effort or glottal damage in performing unsuitable roles.
- In low female voices it's necessary always to evaluate both low and high vocal registers passages.
- There is a precisely accordance between subjective singers' feeling and objective analysis
- There's the possibility and reliability of a predictive evaluation, even if without knowledge of subjective feeling
- We can find a relationship between the singers' medical history, like any vocal desease (even if unknown) or phonosurgery, and analysis results.
It's very important to underline that a singer with unsuited dynamic agility to a specific role can all the same excel in performing it, even if with a higher vocal cost.

REFERENCES

[1] F.Fussi: Fonetografia e tessiturografia nella valutazione della voce artistica. Acta.Phon.Lat., vol.XII, n03, 1990
[2] N.P. Paolillo, M.Chioatto, F.Fussi: Applicabilità del P/P rate al musical; La voce del cantante vol VI, Edizioni Omega, 2010
[3] F. Fussi, M. Gilardone, N.P. Paolillo: Il Vocal Score Profile e il rapporto Partiturografia/Fonetografia: La voce del cantante vol III, Edizioni Omega, Torino 2005.

VOCAL DOSIMETRY (APM) IN OPERA AND MUSICAL SOLOIST SINGERS DURING LIVE PERFORMANCES IN THEATRES: A PILOT STUDY

[1]Nico Paolo Paolillo, [2]Franco Fussi

1 ENT Department Mandic Hospital (Merate-LC) and Teatro alla Scala, Milan, Italy

2 Centro Foniatrico USL Ravenna and Teatro Comunale di Bologna, Ravenna, Italy

Abstract: **In clinical management of singers it's important to identify risk components of vocal fatigue and glottic damage, predicting and assessing the vocal cost of the various vocal performances to avoid any risk. Despite many difficulties, we made dosimetries on singers (9 opera and 9 musical) during live theatre performances.**

Our aim was to evaluate phonatory behaviours before, during and after performances to determine the actual amount of vocal load and the possibility of assessing vocal fatigue and performative potential risks through the identification of a vocal recovery index (VRI). Since the analysis of numerical data from APM (Fo, SPL, vocal doses) doesn't immediately highlight the extent of vocal load, we decided to elaborate data in order to propose a new index: VRI. We found that a lower VRI corresponds to difficult or fatiguing moments, while the opposite happens in moments of vocal rest or recovery. We suppose that there are different threshold ranges between males and females and between different vocal classes, therefore it would be desirable to establish VRI thresholds for references in evaluation of data from dosimetries.

In this study we also show differences between opera and musical soloist singers and describe Fo histogram like a real vocal score profile, SPL histogram like an on stage relative dynamic agility and phonation density graph like an on stage phonetogram to point out many vocal features. Finally, through this method, it could be possible to adapt technical and behavioural measures to avoid and reduce the risk of vocal fatigue or damage.

Keywords : vocal dosimetry, vocal recovery index, vocal effort, vocal doses.

I. INTRODUCTION

In the clinical management of the artistic voice is important to identify all the risk components of vocal fatigue or glottic damage.

The choice of repertoires unsuited to technical and vocal features, inadequate work planning, phonatory behaviours tending to hyperkinesis with very high total phonation times, due to amount of voicing during rehearsals, performances, breaks, teaching, private life, a lifestyle that does not include a regular diet, regular sleep-wake cycles, use of drugs or doping substances, the environment in which it takes place the voice activity are all factors that increase the risk of performing complications. For this reason it is essential to predict and to assess the vocal cost of the various vocal performances to avoid any risk of glottic damage.

II. METHODS

Initially we conducted preliminary clinical assessments (anamnesis, tonal audiometry and videolaringostroboscopy to assess singers' vocal health) and interviews to describe singers' feelings about sung roles, by examining different points of the vocal score, to identify difficult or fatiguing moments and rest times. Later we started to make dosimetries on opera and musical singers during live performances in theatres, using the APM model 3200 on 19 singers, 10 opera and 9 musical singers.

The Ambulatory Phonation Monitor (APM) is a portable, wearable device for objectively documenting the key phonatory behaviors of a client over a full day of normal vocal activity. Specifically, APM measures the amount of time a client phonated, when the phonation occurred, and estimates the client's vocal intensity (dB SPL) and fundamental frequency (F0) during all phonatory activity. This data can be viewed graphically and quantitatively through APM software. The data essentially provides a "profile" of a client's "typical" phonatory behaviors. During the period of monitoring. APM does not record the client's speaking or singing, it only extracts phonation related parameters. [1,2]. The main parameters are: Phonation Time (total duration of phonation expressed as the total cumulated time and the percentage of time spent phonating for the time period of the displayed), Fundamental frequency, mode (values at which most phonation occurred in displayed data) and average [2], Sound Pressure Level and Vocal Doses (derived from mathematical processing of previous parameters).[1] These are the Total Cycles of Vibrations (*Dc*: total number of glottal cycles detected in displayed data) and the Total Distance Dose (*Dd*: estimate of "how far" vocal folds traveled in displayed data in meters).[3]

Analyzing numerical data of frequency, amplitude and vocal dose we don't immediately point out the extent of vocal load. It was therefore decided to assess the ratio between cycles of vibration dose (Dc) and total distance dose (Dd) in order to determine the average vibration

cycles made for each meter of distance traveled in the displayed examination.[4]

III. RESULTS

In this paper it's shown the example of two experienced tenors who have performed at Teatro alla Scala in Milan two different roles in two operas by Giuseppe Verdi (Radames in "Aida" and Jacopo in "I due Foscari") and the example of two female musical singers performing different roles in the musical Cats by Andrew Lloyd Webber. The first example shows the examination of Radames tenor: the total length of the representation in 3 hours and 36 minutes for a PT about 44 min. For the total duration of the opera the ratio between the two parameters (Dc/Dd) is 54 cycles per meter, i.e. 1.54 cm for a cycle. (Fig.1)

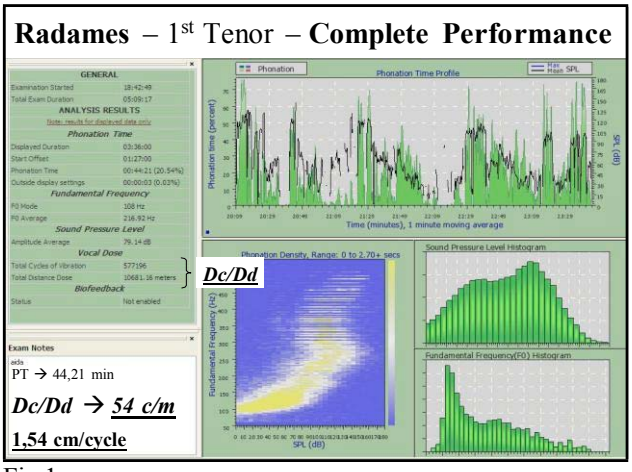


Fig.1

During the interview, the singer has revealed that the most difficult moment of the entire opera is the beginning of the first act, corresponding to the aria "Celeste Aida". By analyzing the parameters regarding only the execution of the aria, we get the result of a ratio Dc/Dd down to 30.2 cycles per meter, that is 3.3 cm cycle: a clear reduction in this ratio; moreover the values of F0 and average amplitude are much higher. (Fig.2).

Analyzing the time of the break between 2nd and 3rd Act and non singing part during beginning of 3rd Act, when it is assumed that there is no phonatory fatigue, but tendency to vocal recovery, we note a marked reduction of Average F0 and amplitude and a ratio Dc / Dd of 208 cycles per meter, idest 0.48 cm cycle. From phonation density graph we can point out speech and soft phonation trend. (Fig.3)

In summary we see that the moments considered the most difficult and fatiguing by the singer are characterized by a lower ratio Dc/Dd and a higher distance in centimeters travelled for each cycle; the opposite happens in moments of vocal rest or recovery. (Tab.1). Similar results for the second tenor: warming up value of ratio Dc/Dd 107 c/m and most difficult and fatiguing moment, the aria in first act , 60.5 c/m. We also analyze the examples of two female musical singers, doing the same considerations.

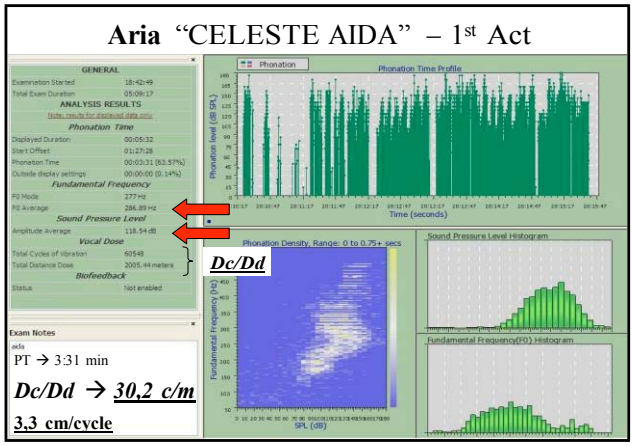


Fig.2

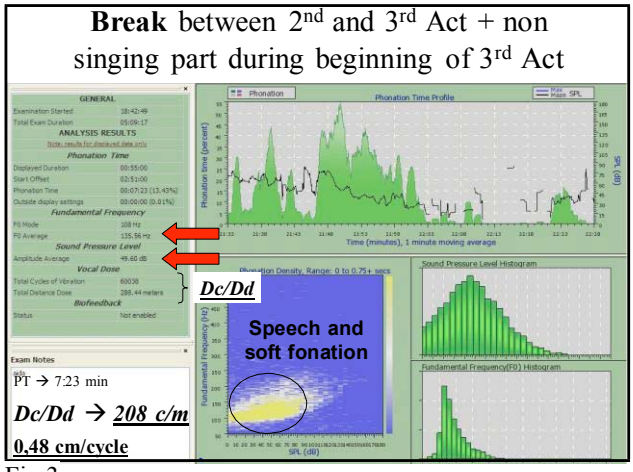


Fig.3

Table 1

VRI (Dc/Dd) in different moments of the performance (Aida Tenor)

• Dc/Dd complete performance	→ 54 c/m	→ 1,54 cm/cycle
• Dc/Dd warm-up	→ 53,3 c/m	→ 0,98 cm/cycle
• Dc/Dd **Aria "Celeste Aida"**	→ **30,2 c/m**	→ **3,3** cm/cycle
• Dc/Dd **Break**	→ **208 c/m**	→ **0,48** cm/cycle
• Dc/Dd 3rd and 4th Act	→ 82,3 c/m	→ 1,2 cm/cycle

The role of Grizabella in the musical Cats, a non-danced role, and the role of Jennytutt-a-poys, a danced role. In the first role we conducted the recordings of double performance and of one night stands. (Tab.2) Dosimetry

in the second singer, instead, shows a value of VRI higher than the previous, sign of probably less vocal fatigue, despite the dance.

Table 2

VRI (Dc/Dd) in different moments of the performance (female musical singer)

- RI double performance → 129 c/m → 0,77 cm/cycle
- RI 1st performance → 123 c/m → 0,8 cm/cycle
- RI 2nd performance → 123 c/m → 0,8 cm/cycle
- RI break → **220 c/m** → **0,45 cm/cycle**
- RI one night stand → **110 c/m** → **0,9 cm/cycle**

IV. Discussion

Making comparisons between the two tenors we found that the phonation time profile of the warm-up and first act shows that the profile of the first tenor is broken, a sign that he tends to rest and needs less warming-up than the latter, that instead works more with increased risks of voice fatigue. (Fig.4)

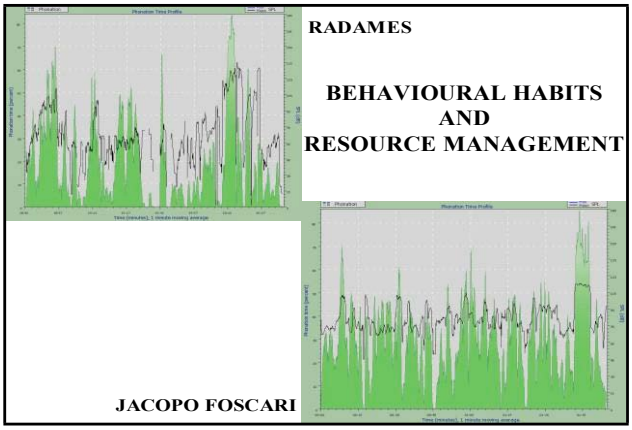


Fig.4

The Fundamental frequency histogram could be considered such as a real vocal score profile. Here we see a substantial equality between the two tenors, a sign that the roles of the two operas have the same musical characteristics in the first act [5,6] (Fig.5) as also evidenced by the histogram of the vocal score profile. [8,9] (Fig.6). The SPL histogram, however, could be considered as such as an on stage relative dynamic agility. In this case the first tenor has a wider range of SPL compared to the second tenor, which uses high SPL for longer, a sign of higher vocal fatigue. [1,6] (Fig.7) The phonation density graph is like a phonetogram and can point out the vocal, technical and behavioral features. The first tenor has a wider phonetograpic range in all vocal sectors and a wider range in speech and soft fonation sector than the latter. A 2006 study [7] shows that inability to produce soft phonation increases when vocal effort is present. In this case the chart can show a very small range of speech and soft phonation in the second tenor, even this sign of an increased vocal effort. (Fig.8)

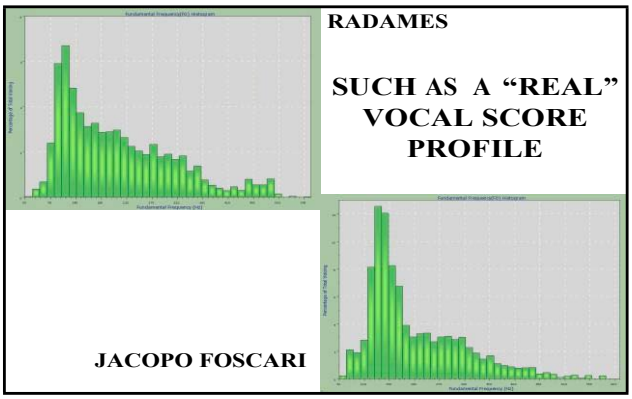


Fig.5

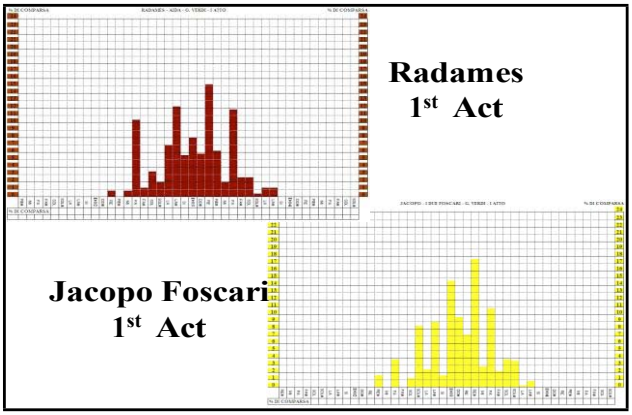


Fig.6

Our aim was to evaluate phonatory behavior before, during and after the performances to determine the actual amount of vocal load and therefore the possibility to assess vocal fatigue and performative potential risks through the identification of a vocal fatigue or vocal recovery index.

In recovery or rest periods the Dc/Dd ratio (VRI) increases, while it decreases during fatiguing periods. Certainly there are different threshold ranges between males and females and between different vocal classes, but we need more studies to explore this aspect. After all that we can consider the ratio Dc/Dd as a recovery index? Another consideration: the more voicing is fatiguing, the more distance for 1 cycle lengthens, thus indicating an hyperkinetic tendency and therefore a greater risk of vocal folds damage when the distance traveled by a vibratory cycle increases. Results regarding the first musical singer previously analyzed (Tab.2) show again a

net increase of VRI in times of less fatigue or vocal rest, while there is a reduction in the most challenging moments, especially in one-night stands than in double performance, perhaps because of awareness to face many hours of work in double performance; this would induce the singer to save energy in order to not tire herself in two consecutive performances.

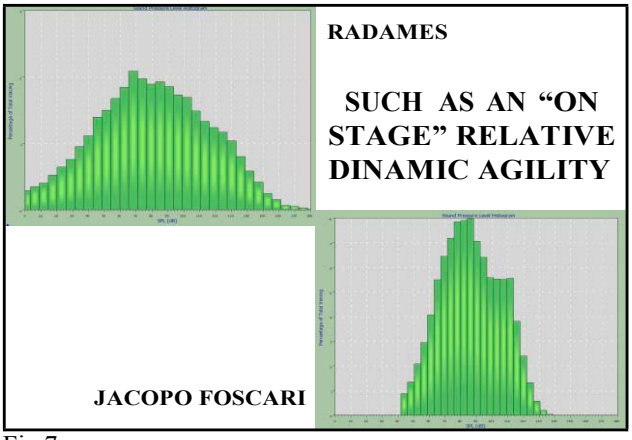


Fig.7

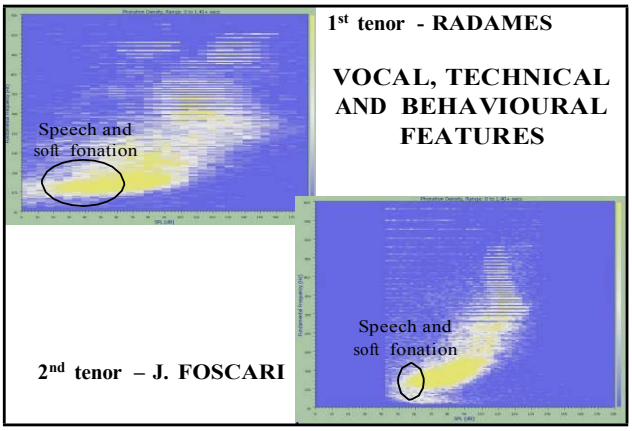


Fig.8

In the second musical singer we found a VRI higher than the previous, sign of probably less vocal fatigue, despite the dance, due to the use of a classical singing technique that allows to sing in a lower average glottic SPL, enhancing use of vocal tract resonators.
According to these data it's possible to make comparisons between musical and classical singers.
Opera soloist singers make 2-3 performances a week and rehearsals with a total sung phonation time less than 5 hours a week. This would allow us to say that there is a greater likelihood of damage from acute fatigue.
Musical soloist singer make 8-9 performances a week, rehearsals and often hard bodily fatigue in dancing with a total sung phonation time more 10 hours a week. In this case there would be a greater likelihood of damage from chronic fatigue.

V. Conclusion

The literature is still poor and so there are not many references. In the future it would be desirable to establish vocal fatigue or vocal recovery thresholds for a reference in the evaluation of data from the dosimetries.
We had some difficulties: the size of APM is too large and cause discomfort to the dance and stage movements. The costumes are often too tight and it's impossible to hide the tool. The singers often don't feel safe to go on stage with discomfort and then refuse to wear it. We must clarify how to perform initial calibration in case of those singers who use both classical and modern way to sing during the same performance.
Through this method it could be possible to adapt technical and behavioral measures to avoid or reduce the risk of vocal fatigue or damage.
We must still understand whether the technique and voice features influence vocal dose parameters; for this would be useful to implement the dosimetry for the same role in different singers. We need more studies to establish standard fatigue and recovery thresholds and ranges and to assess the possible differences between males and females and between different vocal classes

REFERENCES

[1] Svec, J., I. Titze, and P. Popolo. "Estimation of sound pressure levels of voiced speech from skin vibration of the neck." J Acoust Soc Am, Vol. 117, pp. 1386-1394, 2005.
[2] Svec, J., P. Popolo, and I. Titze. "Measurement of vocal doses in speech: experimental procedure and signal processing." Logoped Phoniatr Vocol, Vol. 28, pp. 181-192, 2003.
[3] Titze I., J. Svec, and P. Popolo. "Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues." J Speech Lang Hear Res, Vol. 46, pp. 919-932, 2003
[4] Paolillo N.P. "The vocal score profile (VSP)/vocal range profile (VRP) ratio (P/P ratio) and APM in artistic voice evaluation: application tested on opera and musical singers" – "Voice 2010 – The 4th World Voice Congress – Multidisciplinary Voice Community in Harmony" Organized by World Voice Consortium – Seoul, 6-9/09/2010.
[5] Titze I., E. Hunter, and J. Svec. "Voicing and silence periods in daily and weekly vocalizations of teachers." J Acoust Soc Am, Vol. 121, pp. 469-478, 2007.
[6] Paolillo N.P. "La misurazione del costo vocale e del rischio performativo nel cantante", "La voce del cantante" – Vol. 6, Ed. Omega 2010.
[7] Carroll, T., J. Nix, E. Hunter, K. Emerich, I. Titze, and M. Abaza. "Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study." Otolaryngology-Head and Neck Surgery, Vol. 135, pp. 595-602, 2006.
[8] Paolillo N.P., F. Fussi, M. Gilardone.La valutazione fonetografica dell'artista vocale in rapporto alla valutazione partiturografica dei ruoli cantati, ICare 2008, 9: 92-97.
[9] Paolillo N.P., F. Fussi, M. Chioatto. "Partiturografia e fonetografia, elementi di valutazione comparata nella voce artistica: applicazione al musical", "La voce del cantante" – Vol. 5, Cap. 28, 351-372, Ed. Omega 2009.

INFLUENCE OF HORMONE REPLACEMENT THERAPY ON THE SINGING VOICE TESSITURA OF MENOPAUSAL WOMEN

Janaína Mendes-Laureano[1*]; Marcos Felipe Silva de Sá[2]; Juan Ignacio Godino-Llorente[1]; Nicolás Sáenz-Lechón[1]; Víctor Osma-Ruiz[1]; Juana Gutiérrez-Arriola[1]

1. Department Ingeniería de Circuitos y Sistemas, Universidad Politécnica de Madrid, Spain.
2. Department of Gynecology and Obstetrics, Universidade de São Paulo, Brazil.
*Corresponding author

Abstract: **The differences in the functioning of phonatory organs define the divergent aspects between the speaking voice and the singing voice. The human voice represents the most natural and oldest sound source with which music can be reproduced. Physiologically, singing involves the ability to coordinate and control the musculature and structure of the vocal organ. Along life, the human body goes through anatomo-physiological changes, several of them due to the action of sex steroid hormones. The fluctuations of these hormones starting in puberty affect the vocal folds and the larynx. The literature suggests that hormone replacement therapy (HRT) has a beneficial effect on the larynx preveting the voice changes associated with menopause and increasing vocal longevity. The objective of the present study was to compare the vocal tessitura of menopausal female choir singers using HRT or not in order to determine whether hormone replacement interferes with this vocal parameter. The sample consisted of 38 menopausal women divided into four groups in accordance with the hormonal status. No significant difference in singing voice tessitura were detected between female choir singers taking or not HT with estradiol, tibolone and phytohormone.**

Keywords: Menopause, Voice, Choral, Singers, Sexual hormones.

I. INTRODUCTION

The human voice represents the most natural and oldest sound source with which music can be reproduced[1,2]. Physiologically, singing involves the ability to coordinate and control the musculature and structure of the vocal organ in order to produce modulations in the voice, with sounds varying over a wide gamut of frequencies in harmony and melody[3-5]. The demands made by the singing voice of the organ during singing regarding articulation, respiration and phonation are measurably different from those made by normal speech[6]. Singing requires refined muscle adjustment, reflecting a greater difficulty, since the singing voice makes more demands on the phonatory mechanism[3]. Singers represent a unique population in terms of vocal demands, which differ from those of non-singers, being characterized by greater sensitivity to small vocal changes[7,8]. Along life, the human body goes through anatomo-physiological changes, several of them due to the action of sex steroid hormones. The fluctuations of these hormones starting in puberty affect the vocal folds and the larynx[9]. During the menopausal period, an earlier alteration occurs in women, and more markedly so in the singing voice[10]. The literature suggests that hormone replacement therapy (HRT) has a beneficial effect on the larynx[11,12] preveting the voice changes associated with menopause and increasing vocal longevity[13]. The objective of the present study was to compare the vocal tessitura of menopausal female choir singers using HRT or not in order to determine whether hormone replacement interferes with this vocal parameter.

II. METHODOS

The sample consisted of 38 menopausal women divided into four groups:

1. Menopausal group with no hormone therapy (GNHT): 22 women aged 45 to 60 years, mean age 56 years, in menopause for at least 2 years and taking no hormone therapy (HT).
2. Menopausal group consisting of women taking estradiol hormone therapy (GHTE): 6 women aged 45 to 60 years, mean age 57 years, in menopause for at least 2 years and taking estradiol HT by the oral route for at least 6 months.
3. Menopasual groups taking tibolone hormone therapy (GHTT): 5 women aged 45 to 60 years, mean age 57

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

years, in menopause for at least 2 years and taking tibolone therapy by the oral route for at least 6 months.
4. Menopausal group taking phytohormone therapy (GHTP): 5 women aged 45 to 60 years, mean age 57 years, in menopause for at least 2 years and taking phytohormone therapy by the oral route for at least 6 months. All selected volunteers were submitted to otorhinolaryngologic evaluation in order to rule out any lesions in the larynx and/or vocal folds. The vocal tessitura profile was obtained manually using a Roland keyboard tuned in A 2 to 440 Hz, played by a female music teacher and choir conductor with experience in the area. A well-known Brazilian folk song which required variations in frequency from the most grave to the most acute was selected. After explaining the procedure, the conductor started to play the music, guiding the tuning of the volunteer while testing the maximum and minimum frequency achieved. The values of each musical note and their corresponding frequency in Hz were written down manually[14]. Data were analyzed statistically by the Mann-Whitney tests using the GraphPad Prism® software, with the level of significance set at 5%.

III. Results

No significant difference in F2 values were observed when GNHT was compared to GHTE (p=0.5566), GHTT (p=0.9751) and GHTP (p=0.4727). Again, no significant difference in F1 values were observed between GNHT and GHTE (p=0.2510), GHTT (p=0.3991) and GHTP (p=0.0751). When the F2-F1 values were compared between GNHT and GHTE (p=0.1702), GHTT (p=0.3995) and GHTP (p=0.9751), no significant differences were observed. Table 1 presents tessitura based on the mean, standard deviation (SD) and median values of each maximum and minimum musical note in Hz.

Table 1 – Mean (± SD) and median tessitura of the maximum (F2) and minimum (F1) frequencies in Hertz and difference between the F2 and F1 frequencies of the groups taking no hormone therapy (GNHT), taking hormone therapy with estradiol (GHTE), hormone therapy with tibolone (GHTT), and hormone therapy with phytohormone (GHTP).

	F2	F1	F2-F1
GNHT	387.52 Hz (±116.48)	175.43 Hz (±33.54)	212.09 Hz (±116.64)
	349.23 Hz	174.61 Hz	196.00 Hz
GHTE	426.78 Hz (±154.90)	138.61 Hz (±52.45)	288.17 Hz (±139.08)
	397.90 Hz	156.61 Hz	285.28 Hz
GHTT	367.40 Hz (±100.78)	187.00 Hz (±40.89)	180.40 Hz (±140.86)
	311.13 Hz	207.65 Hz	91.13 Hz
GHTP	413.42 Hz (±55.27)	197.84 Hz (±22.73)	208.96 Hz (±58.45)
	392.00 Hz	200.00 Hz	217.39 Hz

IV. Discussion

The relationship between voice and circulating sex steroid hormone levels has been previously established[15-17]. Several studies have attempted to prove a correlation between menopause and voice using perceptive and acoustic evaluations. Some of these studies have proven this correlation[9,18-21] while others have not [22-26]. The methodology used in the cited studies was perceptive and/or acoustic analysis of the phonation of a sustained vowel or sentence and therefore the measurements were not made during singing or the emission of different musical notes.
Some studies[19,27] have stated that HT has a beneficial effect on the larynx[11,12], preventing the changes of voice associated with menopause and increasing vocal longevity[13]. The data of the present investigation agree with other studies[22,26] that evaluated various vocal parameters and did not detect a significant difference between menopausal women using HT or not.
In the present study, the tessitura of the singing voice of the group taking no HT did not differ significantly from that of HT users. There are no studies in the literature on the normal tessitura values of singing voice in menopausal women, nor any comparisons of tessitura between menopausal women taking HT or not. Thus, the present study contributed to the literature by providing the tessitura values of singing voice of menopausal women taking HT or not.

V. Conclusion

No significant difference in singing voice tessitura were detected between female choir singers taking or not HT with estradiol, tibolone and phytohormone.

References

[1] Camargo TF, Barbosa DA, Teles LCS. Características da fonetografia em coristas de diferentes classificações vocais. Rev Soc Bras Fonoaudiol. 2007;12(1):10-7

[2] Tepe ES, Deutsch ES, Sampson Q, Lawless S, Reilly JS, Sataloff RT. A pilot survey of vocal health in young singers. J Voice 2002 Jun;16(2):244-50.

[3] Ribeiro LR, Hanayama EM. Perfil vocal de coralistas amadores. Rev CEFAC 2005 Abr-Jun; 7(2): 252-66.

[4] Sulter AM, Schutte HK, Miller DG. Differences in phonetogram features between male and female subjects with and without vocal training. J Voice 1995 Dec;9(4):363-77.

[5] Watts C, Barnes-Burroughs K, Estis J, Blanton D. The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. J Voice 2006 Mar;20(1):82-8.

[6] Braun-Janzen C, Zeine L. Singers' interest and knowledge levels of vocal function and dysfunction: survey findings. J Voice 2008 Jul: 23(4):470-83.

[7] Cohen SM, Jacobson BH, Garrett CG, Noordzij JP, Stewart MG, Attia A, et al. Creation and validation of the singing voice handicap index. Ann Otol Rhinol Laryngol. 2007 Jun;116(6):402-6.

[8] Rosen CA, Murry T. Voice handicap index in singers. J Voice 2000 Sep;14(3):370-7.

[9] Amir O, Biron-Shental T. The impact of hormonal fluctuations on female vocal folds. Curr Opin Otolaryngol Head Neck Surg 2004 Jun;12(3):180-4.

[10] Rocha TF, Amaral FP, Hanayama EM. Extensão vocal de idosos coralistas e não coralistas. Rev CEFAC 2007 Abr-Jun; 9(2): 248-54.

[11] Sataloff RT, Lawrence L, Hawkshaw MJ. Medications and their effects on the voice. In: Benninger M, Jacobson B, Johnson A. Vocal arts medicine: the care and prevention of professional voice disorders. New York: Thieme Medical Publishers; 1994.

[12] Lindholm P, Vilkman E, Raudaskoski T, Suvanto-Luukkonen E, Kauppila A. The effect of postmenopause and postmenopausal HRT on measured voice values and vocal symptoms. Maturitas. 1997 Sep;28(1):47-53.

[13] Landau C, Cyr MG, Moulton AW. O livro completo da menopausa: guia da boa saúde da mulher. Trad. H Lanari. Rio de Janeiro: José Olympio; 1998.

[14] Pinho SMR, Bastos PRJ. Quadro para avaliação vocal de correspondentes tonais. Barueri: Pró-fono; 2003.

[15] Damsté PH. Voice change in adult women caused by virilizing agents. J Speech Hear Disord. 1967 May;32(2):126-32.

[16] Ferguson BJ, Hudson WR, Mccarty KS. Sex steroid receptor distribution in the human larynx and laryngeal carcinoma. Arch Otolaryngol Head Neck Surg. 1987 Dec;113(12):1311-5.

[17] Newman SR, Butler J, Hammond EH, Gray SD. Preliminary report on hormone receptors in the human vocal fold. J Voice 2000 Mar;14(1):72-81.

[18] Molina KL, Brasolotto AG, Berretin-Felix G, Cristovam LS. Modificação na freqüência fundamental da voz associada a manifestações de tensão pré-menstrual. Revista Fonoaudiologia Brasil 2000; (4): 7-13.

[19] Tonisi GABR. Efeitos do climatério na freqüência fundamental. Rev CEFAC 2000 Jan-Jun; 2(1):73-80.

[20] Meurer EM, Wender MCO, Corleta HE, Capp E. Female suprasegmental speech parameters in reproductive age and postmenopause. Maturitas 2004 May 28;48(1):71-7.

[21] Schneider B, Trotsenburg M, Hanke G, Bigenzahn W, Huber J. Voice impairment and menopause. Menopause. 2004 Mar-Apr;11(2):151-8.

[22] Fernandez RL, Damborenea DT, Rueda PG, Garcia-Garcia E, Leache JP, Campos MAA et al. Acoustic analysis of the normal voice in nonsmoking adults. Acta Otorrinolaringol Esp. 1999 Mar;50(2):134-41.

[23] Meurer EM, Wender MCO, Corleta HE, Capp E. Phono-articulatory variations of women in reproductive age and postmenopausal. J Voice. 2004 Sep;18(3):369-74.

[24] Mendes-Laureano J, Sá MFS, Ferriani RA, Reis RM, Aguiar-Ricz LN, Valera FCP, et al. Comparison of fundamental voice frequency between menopausal women and women at menacme. Maturitas 2006 Sep 20;55(2):195-9.

[25] Mendes-Laureano J, Sá MFS, Ferriani RA, Reis RM, Aguiar-Ricz LN, Valera FCP, et al. Impact of menopause and hormonal replacement therapy on harmonics-to-noise-ratio of the voice. Maturitas 2007 Feb 20;56(2):223-4.

[26] Mendes-Laureano J, Sá MFS, Ferriani RA, Romão GS. Variations of jitter and shimmer among women in menacme and postmenopausal women. *J Voice* 2009; 23: 687-689.

[27] Stoicheff ML. Speaking fundamental frequency characteristics of nonsmoking female adults. J

HOW CAN POSTURO-ACOUSTIC SYSTEM HELP THE SINGER IN VOICE QUALITY RESEARCH?

Joseph Quoidbach

IMEP - Institut Supérieur de Musique et de pédagogie, Namur, Belgium

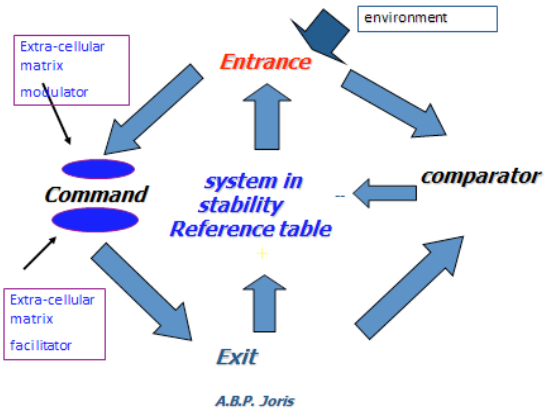
Abstract: **Since September 2010, a Postural-acoustic Lab has been taking shape at the IMEP School of Music and Music Education in Namur, Belgium.**
Thanks to recent developments about the Postural System in the Field of Neurophysiology , as well as to progress in Information Technology and Robotics, it is now possible to provide Music Students with a modern set of tools leading to the optimizing of an ergonomic position in performing, regardless of the instrument.
It's about moving away from the sole system of "retro-control" to a system of anticipation, in other words, from a system of feedback to a system of "feedforward".

INTRODUCTION

In this paper the following points will be addressed:
-What are the benefits of the proposed approach in which three means of "retro-control" are applied simultaneously, as part of the musician's strategy to verify posture as sound is emitted.
-The notion of posture and posturology
-Postural-acoustic laboratory developed at IMEP, Namur
-Conclusions

METHODS

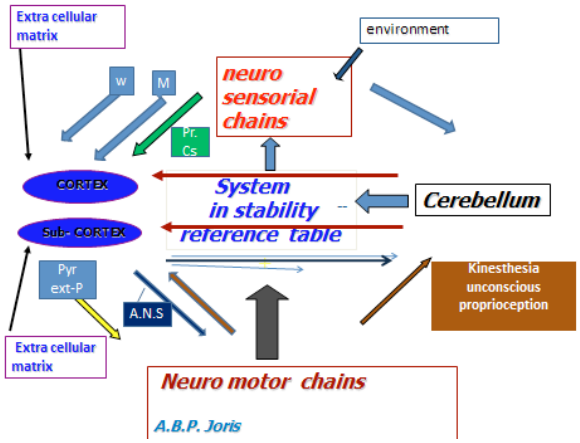
What are the benefits of the proposed approach in which three means of "retro-control" are applied simultaneously, as part of the musician's strategy to verify posture as sound is emitted?
There are a number of them.
To begin with, this approach is about creating the conditions in which musicians learn to move from a system of feedback to a system of anticipation ("feedforward"). This is the starting point that will allow musicians to go from a system of long loops, to a system of short loops.
Secondly, this approach will offer a set of tools allowing musicians to gradually grow in independence regarding the ability to check whether the sound that is actually being produced is the sound that they had the intention of producing.
Musicians also can discover on a screen the relation between harmonics and voice quality
This issue is of particular interest for singers, given the fact that in reality they don't hear the actual sound that they are producing.
Initially, this need for singers led to the idea of a postural-acoustic lab. It was then observed that it could be equally applicable to other instruments besides the voice.
The concept of posture and posturology
In order to fully understand how this postural-acoustic approach works, it is indeed necessary to mention a series of notions about posture as well as some elementary notions about the neurophysiology of the postural system.
What would the definition of posture be?
According to Paillard, posture refers to "the body attitude or to the position of the whole set of segments at any precise moment".
Based on this way of referring to posture, we can move on to the definition of posturology:
According to Gagey, Posturology is the study of the geometrical and bio-mechanical organization of different segments of an individual in space and of the regulation process involved. In this sense, it is the sum of neurological mechanisms which allow the balancing of these elements in space during the standing position or during the walking action.
The notion of a postural system implies inputs and outputs, as well as a central computer processing the stream of information.
The term "postural system" obviously includes the notion of "system".
By its own definition, a system is a combination of elements put together in a way that allows them to become a whole.



In general, and in normal conditions, in terms of its referential elements the postural system is in balance.
When taken away from its balance, it refers to inputs, which will in turn give information to a commanding system and a comparer.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

These last two elements will influence outputs whose actions will bring the system back to balance.
When referring to the system concerning the human body, the pattern as illustrated by Doctor Maurice Joris, President of the Belgian Society of Posturology is as follows:
"Posture is based on a notion of stability or balance, in other words, the fact of returning to the initial position after having left it. In order to obtain this stability, it is necessary to use a reliable postural system.
In order to understand the logic behind the postural system, here's a systems analysis.



When the system moves away from its reference points, or from its maneuvering margin, it refers to inputs which are neurosensory chains, which in turn influence the cortex and subcortex through a "three-lane highway".
These three routes are:
-the path of conscious "self-perception"
-the vascular mechanical-receptors described by Mittelsteadt
-the spinal column according to Wike
The neurosensory chains equally influence the cerebellum through specific paths.
The cortex and the subcortex influence the neuromotor chains through the paths of the pyramidal and extrapyramidal system as well as through paths of the autonomous nervous system.
The neuromotor chains give information thanks to the paths of the kinaesthesia (unconscious "self-perception") for the cortex and the subcortex.
Attention should be drawn to the fact that the neurosensory chains in which the eyes, the internal ear, the masticating system, the skin, as well as the mechano-receptors of the vascular system are found, are under the environment's influence.
The extra-cellular matrix, which is either modulator of facilitator, influences the cortex or the sub-cortex.
The whole of these mechanisms of great precision should allow individuals to place themselves correctly in space, and to perceive their subjective vertical line as well as their physical vertical line.
In reality, this process should happen quite naturally in most cases, but unfortunately, we are forced to acknowledge that there is often divergence. The process of integrating the physical vertical line often takes place inaccurately. People think that they are in a certain position while they really are in another.
There is often a sensory conflict.
This is the reason why it seems important to establish an approach offering a set of tools allowing people to correct these inaccuracies.
For musicians it is imperative to be able to considerably adjust their position, in order to create the conditions that allow the production of the most performing sound, as close as possible to the reality of sound they aim for.
Postural-acoustic Laboratory developed at the IMEP (School of Music) in Namur, Belgium.
A Postural-acoustic Laboratory has been taking shape since 2010 at IMEP (School of Music and Music Education) in Namur, Belgium.
Since 2003, with the aim of constantly improving the quality of teaching, experiments in the field of posture and acoustics began to take place. Unfortunately the Information Technology potential at the time didn't allow for confirming or questioning the hypotheses being put forward through our approach.
The core of the subject was to prove that: body posture has an influence on the quality of sound production, regardless of the instrument being played.
Progress in Information Technology has allowed the way of functioning to evolve in a very positive way.
In June 2009, a physiotherapy student came with the request of some assistance in her research for the end of her undergraduate studies.
In June 2010, she presented the results of her work and research which had taken shape thanks to the resources (both technological and human) made available and put at her disposal by the IMEP.
The subject of her research (end of studies project) was "The influence of Posture on Sound Quality in Students majoring in Voice Studies".
She managed to demonstrate that there was in fact a very close correlation between body posture and the quality of sound production. Based on these results, it became obvious that it was important for voice students to have access to a reliable and tangible set of tools.
An article published by Professor Richard Miller and Juan Carlos Franco in the National Association of Voice Teachers' Journal , followed in June 1995 by the Voice Teachers Association Bulletin already spoke about the "Spectography" of the singing voice.
We became interested in this publication, and based, among other sources, on this particular one, we have shaped and brought together a series of elements put into practice in our School of Music, which are the object of our conference today.
What are the functions of this approach or set of tools?
-To establish a sound "identity card" from the time of enrollment to the school. Students can record

themselves and become aware of the physical points of reference of the produced sounds.
-To allow a linear follow-up of the way in which the sound is evolving throughout the five-year program. A computerized file gives students and teachers the opportunity to visualize the specific and objective physical characteristics involved in the production of sound, and how they are being transformed as this awareness grows through the application of the proposed approach.
-To make use of a feedback procedure.
Musicians are able to observe, either directly or in a re-play of a recording, on screen, the various significant curves of the produced sound.
-To pursue a body attitude which is in correlation with the sound, by means of a feed-back procedure which is made possible in a visual manner through the use of large mirrors placed at a 45° angle.
Musicians can see themselves simultaneously from the front and from the side, which facilitates the possibility of seeking the most suitable and ergonomic position, depending on the instrument being played.

What are the elements needed for this postural-acoustic laboratory?

- A set of standing mirrors, placed at a 45° angle
- A force platform, such as the "biorescue" type, allowing the musician to measure a series of points of reference, mostly foot tracks on the ground, and to control the center of pressure in the standing and still position, as well as while singing.
- A microphone (Neumann type) linked to a sound card and to an Audio spectrographical analyzer. (City)

Working session procedure:
Take the case of a singer. We ask him/her to stand on the platform, to check his points of pressure, to control verticality thanks to the mirrors, then to begin to sing in the required position.
It's possible to vary the points of reference, as follows:
-head position backwards and forwards, left/right rotation of the cervical column as well as left/right inclination of the cervical column.
-position of the tongue
-Transfer of the center of pressure of the body, in various directions.
-With a sound in the medium register as a starting point, we explore the other registers.
-The spectogram will be evaluated according to the preceding combinations.
-It's crucial to evaluate how the basic overtones as well as the multiple overtones evolve.
-The singer eventually recognizes the position in which the sound production will be at its best. By repeating the result again and again, he/she will memorize and engram it.

CONCLUSIONS

It's important to provide a School of Music with the most advanced resources now available in order to allow an optimal development of the Students' musical skills. Thanks to the progress in Technology, as well as in the field of Posturology, we now have access to such resources.
We are at the dawn of fascinating work, with an endless scope of research in various fields.

REFERENCES

[1] anatomie clinique du système nerveux central professeur Prades Masson édition
[2] posturologie clinique. Dysfonctions motrices et cognitives Weber et Villeneuve, 2007, éditions Masson.
[3] Le sens du mouvement Alain Berthoz éditions Odile Jacob
[4] posturologie clinique tonus, posture et attitude, 26e journée de posturologie clinique. B. Weber et Ph. Villeneuve 2009.
[5] Analyse spectrographie de la voix chantée. Richard Miller et Jaun Carlos Franco in N A TS journal septembre 1991. (Journal de l'association nationale des professeurs de chant aux États-Unis).

Session IV:
Signal analysis

ESTIMATION OF MULTIPLE SOURCE COMPONENT USING GENETIC ALGORITHM

Cheolwoo Jo[1], Jaehee Kim [1]

[1] School of Mechatronics, Changwon National University, Changwon 641-773, Korea

Abstract: **Source of speech signal consist of voiced part and unvoiced part. In conventional source-filter model, those two sources are considered to be independent. But in real situation it is difficult to segregate the source into voiced and unvoiced part. Actual source consist of mixture of two sources and the ratio varies according to the contents or intention of the speaker. In this paper we tried to segregate the components of voiced and unvoiced while considering source models. Source signals are modeled based on residual signal measured from inverse filtering. Two kinds of source models are assumed. Each model parameters are optimized to the original speech signal using genetic algorithm. The resulting parameters were compared in terms of the mel-cepstral distance to the original signal, spectrogram and spectral envelope from the synthesized signal.**

Keywords : **Voice, source, model, synthesis, optimization**

I. INTRODUCTION

Voice source can be utilized in various areas such as speech synthesis, speech recognition, pathological voice processing, speech coding etc. In speech synthesis, for example, voice source is very important because it has big effect on the quality of the synthesized speech in terms of naturalness, intelligibility and emotional expression. In other case, to measure the parameters of the disordered voice, there are many parameters which are related to voice source. There have been many previous researches which tries to measure the source informations from the speech signal [1] [2] [3] [4].

Voice quality can be measured in various ways. The most precise way to observe the vocal folds is biological measurements. But it is not easy and not convenient. So naturally the indirect measurement from acoustic speech signal is preferred. But because of some limits on the mathematical analysis methods, there is no single way to extract voice source parameters. One simple way is to estimate the source component from the numerical analysis.

Multiple source estimation was carried out in previous researches in the area of speech coding and speech synthesis. But their method focused on the approximate estimation of source by frame based analysis method and the purpose was not on finding exact ratio from the specific source model. So our aim for this research is to estimate the two components and obtain the numerical ratio of the two sources from the analysis of speech signal considering specific source model.

II. SOURCE ANALYSIS METHODS

In this paper, we use the inverse filtering from the linear predictive analysis to estimate the voice source. LP(linear prediction) method is a well-known method which models a signal or a system into a form of mathematical function. It is the best method measuring residual signal from LP analysis to estimate the glottal activities [5].

According to the source-filter theory, voice source consist of impulse train, which represents voiced part, and random noise, which represents unvoiced part. In simple source model speech signal is divided into voiced/unvoiced/silence part on temporal basis. Only one kind of the three can be possible in simple model. But in real situation, voiced and unvoiced part cannot be clearly separated. So in mixture source model, two types of source are considered at the same time. Yegnanarayana et.al.[6] used an iterative algorithm to separate the periodic and aperiodic components based on spectral decomposition.

In our research, we used a genetic algorithm to find an optimal level of noise sources in addition the voiced source, which is estimated from the residual signal. And we used a voice source model simulator to analyze the speech signal.

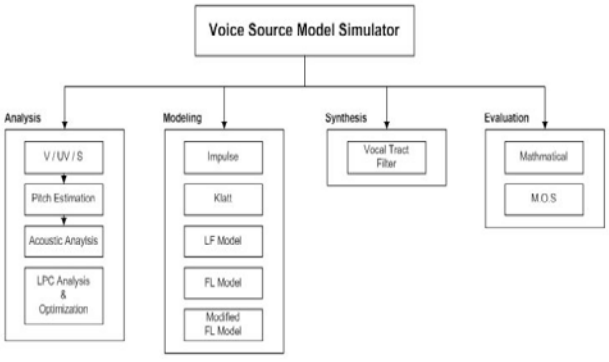


Fig 1. Functions of the simulator

III. VOIVE SOURCE MODEL AND BASIC ANALYSIS PROCEDURE

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

In this research, we considered two types of voice source signal. First one is unipolar source model, the other one is Klatt source model.

Unipolar source model is a simplified impulse model from residual signal. For each pitch period of the residual signal, only the highest peak is chosen as a candidate of the source signal. The remaining pulses are set to zero.

Klatt source model is a model which resembles to the shape of the actual glottal volume velocity. This model corresponds to the integration of the excitation, so this model can be compared to the integrated residual signal.

Analysis of speech is done by conventional linear predictive analysis procedure. Residual signal is the reference source model which can be used to parameterize the voice source.

The residual signal is used to generate approximated source signal based on pitch and amplitude informations of the residual signal. In each excitation position, unipolar and Klatt source shape is located.

IV. Source Optimization

Genetic algorithm was used to find the optimal noise level of the source component. Genetic algorithm uses the random and statistical method to optimize cost function. Table 1 shows options for GA algorithm in this research. Maximum number of iterations are set to 700. These parameters for genetic algorithm are chosen by trial and error method.

Figure 2 shows the flow of the optimization process. Based on original residual signal, noise component ratio is optimized to reduce the error between original speech and re-synthesized speech. On genetic algorithm, algorithm is iterated until the error becomes smaller than pre-specified range.

As a cost function to be minimized, the following functions were used.

For Klatt source model,

$$g(t) = \begin{cases} at^2 - bt^3 , & (0 < t < O_q T_0) \\ 0 , & (O_q T_0 < t < T_0) \end{cases}$$

$$err = |A - g(t)| \quad (1)$$

$$err = |\{A - (D + x(1)C)\}| \quad (2)$$

For unipolor residual,

$$err = |A - (x(1)B + x(2)C)| \quad (3)$$

Where A is the original residual, B is the modified unipolar impulse source model, C is the random noise signal and D is the optimized Klatt source model only voiced part.

Figure 3 shows the process of optimization, one pitch, the Klatt source model used as voiced source model and white random noise used as unvoiced source component.

Table 1. Options for GA algorithm

Options	Values
Population Type	Double Vector
Population Size	200
Creation Function	Uniform
Crossover Function	Scattered
Generation	700
Hybrid Function	fminunc
Mutation Function	Gaussian
Elite Count	5
Stall Generation Limit	100
TolFun	1/inf
TolCon	1/inf

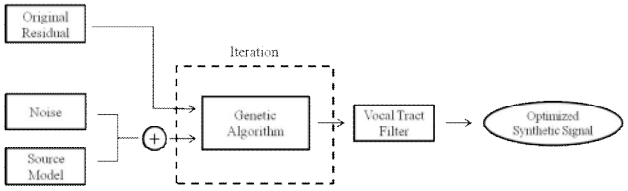


Fig 2. Flow of optimization process

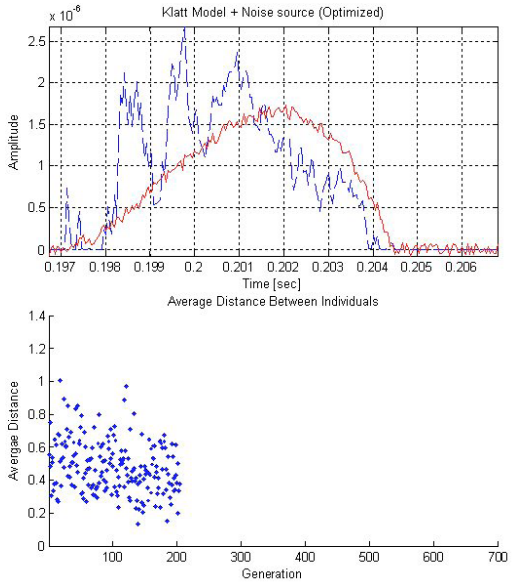


Fig 3. Error minimization from GA

V. Results

Figure 4, 5 and 6 shows examples of comparing the original speech and the resynthesized speech in terms of time and frequency domain. And then figure 7, 8 and 9 shows original signal and the result of the resynthesized signal after optimization in terms of time and frequency domain.

Figure 9 shows us modified unipolar residual and Klatt model give us close similarity in terms of spectral component. But in terms of the melcepstral distance,

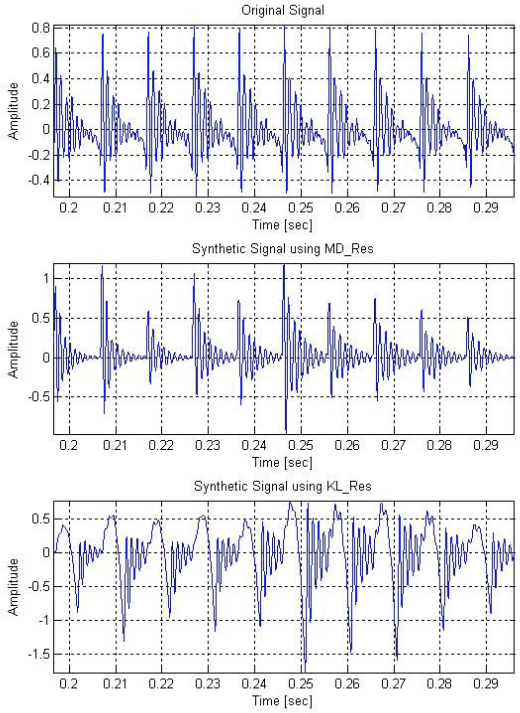


Fig 4. Synthesized signal from each model

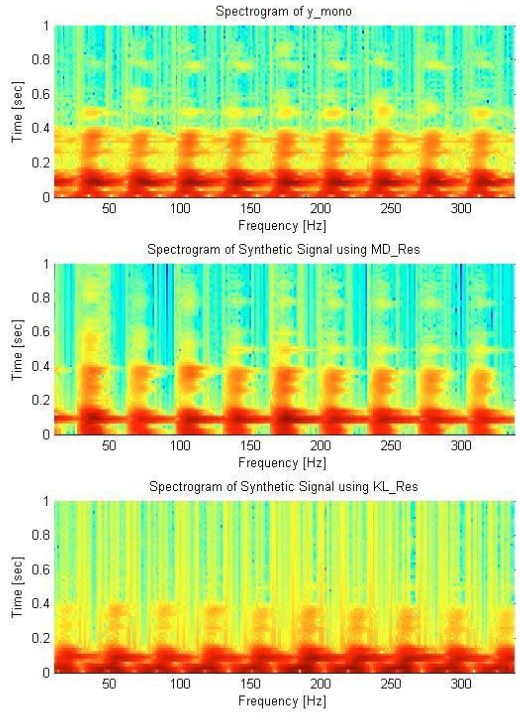


Fig 5. Spectrograms from synthesized signal

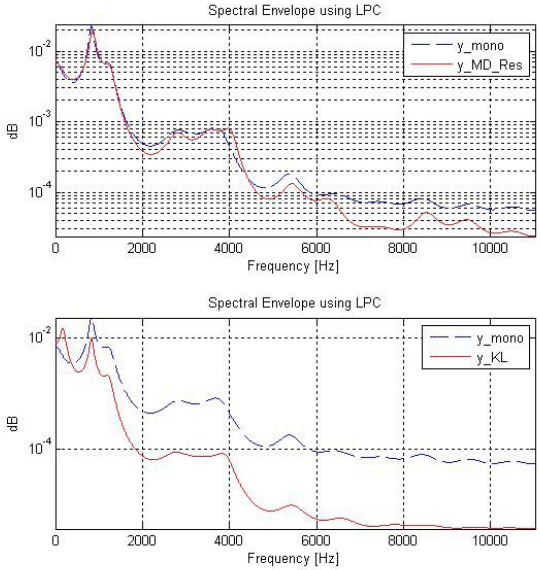


Fig 6. Spectral envelope comparison for non-optimized case

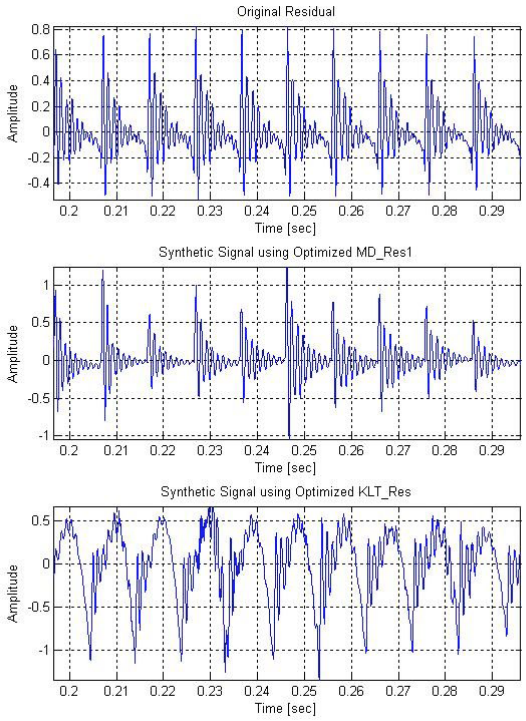


Fig 7. Synthetic signal for optimized source models

optimized version of the modified residual signal showed the closest to the original signal. In case of Klatt source, optimization process reduced the distance the distance considerably and it is useful to estimate noise component of the source in this way.

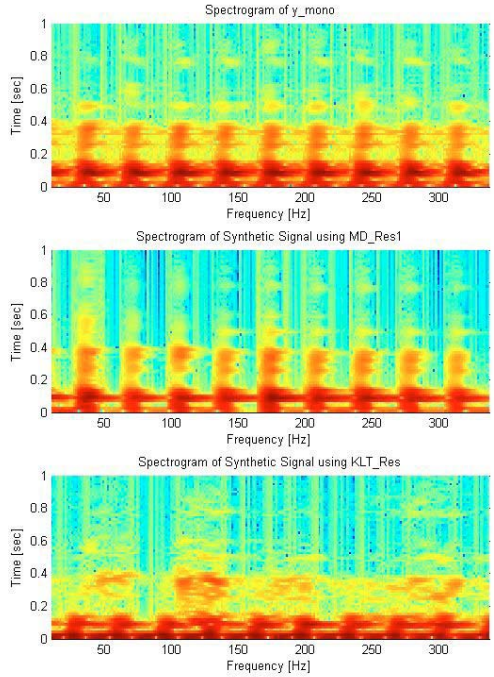


Fig 8. Spectrogram from optimized synthetic signal

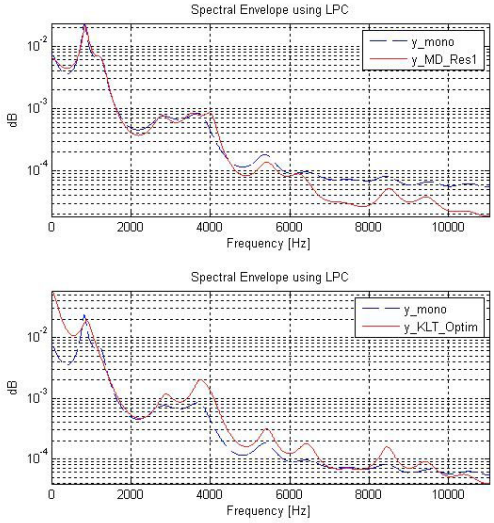


Fig 9. Spectral envelope comparison for optimized case

Table 2. Spectral distance between original speech and each model

Source Model	Distance
Modified Residual(MD_Res)	26.4
MD_Res1 (Optimized)	25.3
Klatt	181.5
Optimized Klatt	148.6

VI. Conclusions

In this paper we tried to estimate voice source components by applying optimization procedure to estimate the voiced and unvoiced components from the speech signal. We used genetic algorithm as an optimization method.

It is found out that addition of noise components with optimization procedure reduced error between the original signal and the synthesized signal when we use voice source models for research purposes. Analysis process for Klatt source model with additional noise with optimization process can be useful for the analysis of speech in various occasions such as speech synthesis or voice quality analysis or pathological voice analysis etc.

In future research, it is required to reduce the spectral distance while adding noise components to the voice source in multiple frequency bands.

Acknowledgement

The authors of this paper were partly supported by the Second Stage of Brain Korea21 Projects.

References

[1] P. Chytil & M. Pavel, Estimation of Vocal Fold Characteristics using a Parametric Source Model, Eleventh Australian International Conference on Speech Science and Technology, Auckland, NewZealand, 2006.

[2] A. Forcin, E. Abberton, Phonetics & measurement of voice quality, VOQUAL'03, Aug. 27-29, Geneva, Switzerland, 2003, pp. 1-27.

[3] P. Mokhtari, H.R. Pfitzinger, C.T Ishi, "Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice-quality, VOQUAL'03, Aug. 27-29, Geneva, Switzerland, 2003, pp. 133-138.

[4] P. Alku, Glottal wave analysis with pitch synchronous interactive adaptive inverse filtering, Speech Communication, 11, 1992, pp. 109-118.

[5] J.D. Markel, Gray, A.H., *Linear Prediction of Speech*, Springer-Verlag, 1976.

[6] B. Yegnanarayana, d'Alessandro Christophe, Darsinos Vassilis, An iterative algorithm for decomposition of speech signals into periodic and aperiodic components, IEEE Trans. on Speech and Audio Processing, Vol.6, No.1, Jan. 1998, pp. 1-11

[7] D.H. Klatt, Software for a Cascade/Parallel formant synthesizer, JASA, Vol. 67, No. 3, March, 1980, pp. 971-994.

IMPACT OF RIGID ENDOSCOPIC LARYNGOSCOPY ON ELECTROGLOTTOGRAPHIC AND ACOUSTIC PARAMETERS

Tobias Meyer[1,2], Jakob Unger[1], F. Peter Schwerdtfeger[2], M. Döllinger[3], Jörg Lohscheller[1]

1 University of Applied Sciences Trier, Germany, Department of Computer Science, Trier, Germany
2 Department of Otolaryngology, Klinikum Mutterhaus der Börromäerinnen, Trier, Germany
3 University Hospital Erlangen, Department of Phoniatrics and Pediatric Audiology, Erlangen, Germany

Abstract: **Rigid high-speed laryngoscopy is the state of the art examination technique for the visualization of vocal fold dynamics. However, due to the insertion of the rigid endoscope in the oral cavity the voice production process including the dynamics of vocal fold vibrations becomes impaired. Therefore, the currently available computerized analysis procedures, which have been designed to enable a highly precise determination of vocal fold vibrations, measure vocal fold dynamics within an non-physiological condition. In this study the influence of rigid laryngoscopy on vocal fold dynamics and on the objectively derived voice measures are quantitatively investigated.**

Keywords : **rigid laryngoscopy, vocal fold, laryngeal imaging**

I. INTRODUCTION

For the clinical examination of voice disorders rigid videostroboscopy is the most widely used examination technique to enable a visual inspection of vocal fold structure and dynamics. With the arise of modern larynx examination techniques such as kymography and high-speed imaging (HSI) a more accurate analysis of the underlying vocal fold vibration pattern became feasible. To derive a precise quantitative analysis diverse post processing procedures have been developed to extract vocal fold dynamics from the image data and quantify the degree of vibration symmetry and regularity [1].

One of these analysis and visualization procedures based on HSI is the computer based phonovibrogramm (PVG). It is able to visualize the entire oscillation pattern separated for each vocal fold within a single image (PVG). It was shown, that a highly precise quantification of vocal fold vibration parameters can be performed [2].

Mainly all results obtained from current analysis approaches base on rigid laryngoscopy. However, during rigid laryngoscopy the formerly undisturbed process of voice production becomes impaired because of at least two reasons: the insertion of the endoscope within the oral cavity and due to the holding of the protruded tongue by the examiner. By protruding the tongue the physiological position of the larynx is changed with the epiglottis into a more anterior superior position and thus putting the larynx with its vocal folds into a higher state of tense [3,4,5].

As a direct consequence quantitative measures, like e.g. PVG, derived from the endoscopic video data and likewise the acoustic signal do not reflect the normal, unimpaired physiological condition anymore.

In this study potential alterations of vocal fold vibrations induced by rigid laryngoscopy were investigated and quantified.

II. METHODS

Forty healthy subjects (20 females and 20 males) with untrained voices and no clinical history of voice disorders were examined during sustained phonation of the vowel /i/ at a comfortable frequency and intensity. The mean age of the female group was 41 (+/-13.1) years and 37.6 (+/-14.4) years for the male group. Each subject was examined twice.

Firstly, to derive information about the unimpaired condition vocal fold dynamics were examined using electroglottography (EGG) and acoustic recordings. EGG provides a relative measure of vocal fold closure without having equipment encumber the oral cavity [6]. EGG and acoustic data were captured using the Laryngograph® (Ltd., London, United Kingdom) system.

Following, a high-speed recording of vocal fold vibrations was performed by rigid laryngoscopy

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

accompanied by a second recording of the EGG and acoustic signal reflecting the impaired voice production condition. The laryngoscopic video and acoustic data were recorded using the Endocam 5562 high-speed camera system, Wolf Corp., Knittlingen, Germany.

To detect alterations of vocal fold dynamics between the different examination situations a set of established parameters like fundamental frequency, jitter, shimmer and normalized noise energy (NNE) were computed from the EGG and acoustic data. Statistical analysis (Mann-Whitney-U-Test) was performed to identify potential significant changes between the different examination setups.

II. RESULTS

The results of the quantitative analysis of the acoustic and EGG data show that objective parameters representing vocal fold dynamics are significantly influenced by the examination situation. Firstly, during rigid endoscopic examination a significant increase of EGG detected fundamental frequency ($p<0.05$) could be identified. Average fundamental frequency was 179.67 Hz during the uninfluenced examination situation and increased to 225.72 Hz during rigid laryngoscopy reflecting a different (higher) muscular tension of the larynx. Fig. 1 further shows the change of the computed EGG-Jitter and acoustic NNE-values derived from the two examination situations. Significantly ($p<0.05$) increased values of Jitter and NNE prove an augmentation of vocal noise during rigid endoscopy. Average EGG-Jitter increased from 0.27 to 0.38%; NNE from -17.19 to -14.37. Accordingly, EGG-shimmer was significantly increased ($p<0.001$) during rigid laryngoscopy, rising from 1.85 to 3.55%. Fig. 2 displays the increase of perturbation measures in the shape of EGG-Shimmer (right) and fundamental frequency (left).

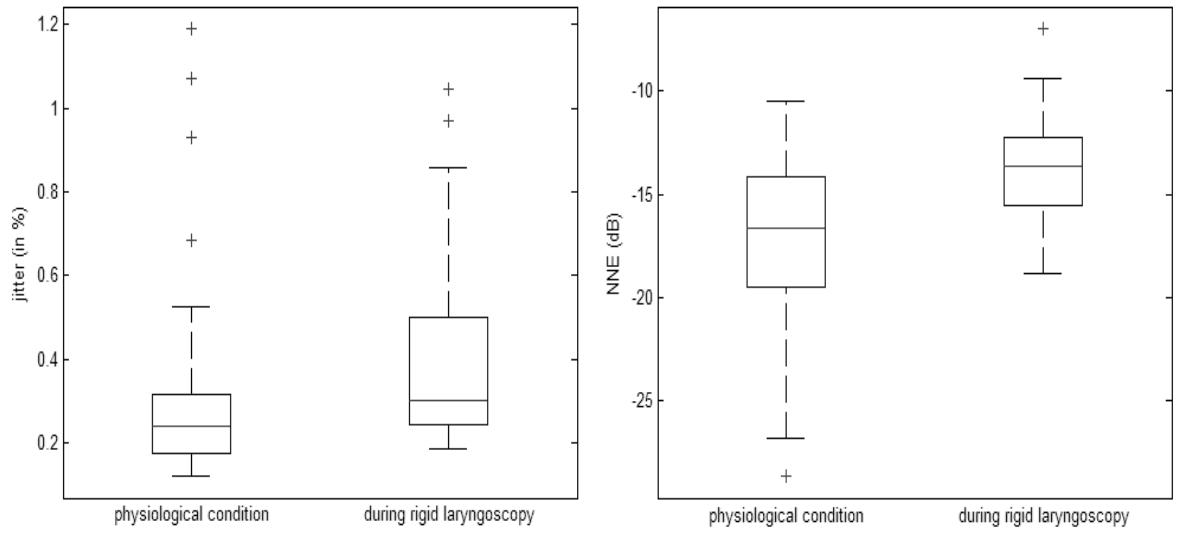


Fig. 1: Left: Boxplots of EGG-Jitter values obtained from the two different examination situations. Right: Boxplots of acoustic NNE values obtained from the two different examination situations. Both parameters are significantly increased as a result of the endoscopic examination.

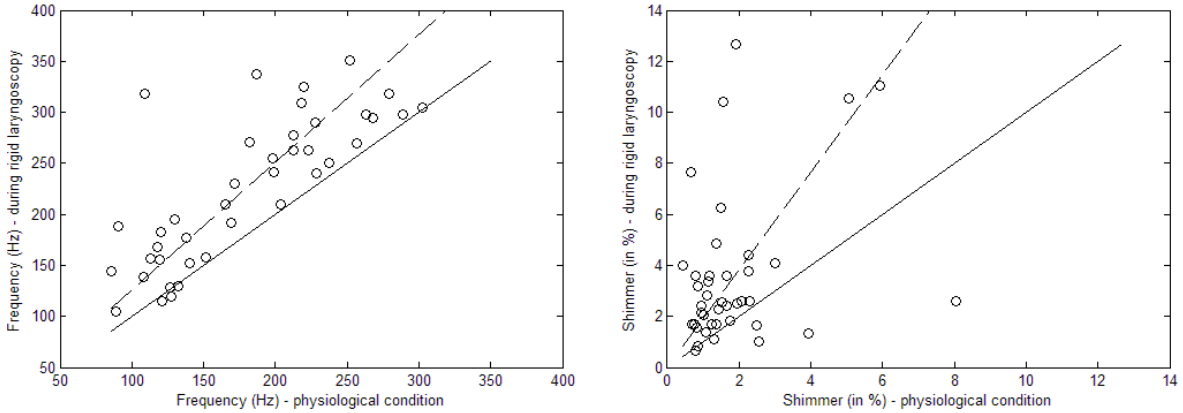


Fig. 2: Data obtained from the two different examination situations. The continuous line represents positions where parameters remain constant in both examinations. The dotted line has minimum distance to all measurements reflecting the averaged linear relationship of parameters from both examination situations. Left: Scatterplot of EGG computed fundamental frequency values. Right: Scatterplot of EGG computed shimmer values. Both parameters are significantly increased as a result of the endoscopic examination.

III. DISCUSSION

The results of the study demonstrate an increased irregularity and thus alteration of vocal fold vibrations induced by rigid laryngoscopy. The vibration pattern of vocal folds is hereby influenced by the inevitable change of the subjects' head position during the examination and by the endoscope within the oral cavity itself. Hence, when applying rigid laryngoscopy it has to be taken into mind that the examination situation itself influences significantly the dynamics of vocal folds. Particularly, when applying modern high-speed imaging systems which facilitate principally more detailed information about vocal fold vibrations it has to be considered that the obtained parameters are likewise altered during the non-physiological examination.

V. CONCLUSION

Objective parameters reflecting vocal fold dynamics and acoustic voice signals are significantly affected by rigid laryngoscopy. Rresults obtained from high-speed videos and rigid endoscopy reflect vocal fold dynamics within a non-physiological state. Hence, normative values about undisturbed vocal fold vibrations are methodically difficult to obtain. However, a combination of flexible endoscopy and high-speed imaging would improve the accuracy of vocal fold analysis procedures.

VI. ACKNOWLEDGGEMENT

This work is supported by the German Research Foundation (DFG), LO-1413/2.

REFERENCES

[1] Deliyski DD, Hillman RE. State of the art laryngeal imaging: research and clinical implications. *Curr Opin Otolaryngol Head Neck Surg.* 2010 Jun;18(3):147-52. Review.

[2] Lohscheller, J., Eysholdt, U., Toy, H., and Döllinger, M. "Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics," *IEEE Trans. Med. Imaging* 2008 Mar; 27, 300–309.

[3] Schade G, Hess M. Flexible versus rigid laryngoscopy and stroboscopy. Differential findings in voice disorders. *HNO*. 2001 Jul; 49(7):562-8. German.

[4] Schröck A, Stuhrmann N, Schade G. [Flexible 'chip-on-the-tip' endoscopy for larynx diagnostics]. *HNO*. 2008 Dec; 56(12):1239-42. German.

[5] Chandran S, Hanna J, Lurie D, Sataloff RT. Differences Between Flexible and Rigid Endoscopy in Assessing the Posterior Glottic Chink. *J Voice*. 2010 Nov 3.

[6] Ronald J. Baken, Electroglottography. *J Voice*. 1992 Vol. 6, No. 2, 98-110No.

AUTOMATIC GRBAS ASSESSMENT USING COMPLEXITY MEASURES AND A MULTICLASS GMM-BASED DETECTOR

J. D. Arias-Londoño[1], J. I. Godino-Llorente [2], N. Sáenz-Lechón[2], V. Osma-Ruiz[2], J.M[a] Gutiérrez-Arriola[2]

[1]Bioinstrumentation Research Group, Universidad Antonio Nariño, Bogotá, Colombia.

[2]Department of Circuits & Systems Engineering, E.U.I.T de Telecomunicación, Universidad Politécnica de Madrid, Spain.

Corresponding author: J.D. Arias-Londoño (julian.arias@uan.edu.co)

Abstract: **This paper presents a system for the automatic assessment of pathological voice quality according to the GRBAS protocol, which uses a short time scheme and a characterization based on 9 complexity measures, including conventional nonlinear statistics and 7 entropy based features. The classification is carried out using three different multiclass classification strategies all of them based on Gaussian Mixture Models. The performance of the system is measured in terms of efficiency and a statistical agreement index. The results show that the complexity analysis provides relevant information for the automatic assessment of voice quality according to the GRBAS protocol.**

Keywords: **Automatic GRBAS assessment, Complexity measures, multiclass classification.**

I. INTRODUCTION

In the clinical environment, evaluation of voice is usually carried out by means of a combination of perceptual evaluations and acoustic parameterizations of the speech trace. Perceptual evaluation consists on a subjective diagnosis of voice quality, based on comparisons with other voices, patients or with previous impressions of the same voice. The main problem is that a reliable perceptual analysis requires a standardized ability to avoid inter and intra listener differences in the evaluations [1]. Although the assessment based on acoustic parameters is becoming a usual technique of analysis, perceptual evaluation is still the most practiced method for the evaluation and clinical management of voice disorders [2]. Unfortunately, a good correlation between acoustic parameters and perceptual evaluation of voices remains unfound [3].

Perceptual evaluation has been widely criticized because its subjectivity. As a result, the reliability of the evaluation is not always adequate and auditory perceptual ratings can be confounded by factors such as the listener's perceptual bias, the listener's experience, the type of rating scale used, the listener's fatigue, the perceptual sensitivity of the listener to a particular voice feature and to the voice sample being evaluated [4]. This situation can be improved using an automatic system, which should provide accurate, reproducible and graded measures of a patient's voice quality, helping speech and language therapists with the patient's treatment and rehabilitation [5]. However, few efforts have been performed in this way due to lack of standardized protocols and also low correlation with objective acoustical analysis. Currently, the most widely accepted and recommend by The Japanese Society of Logopeadics and Phoniatrics and the European Research Group evaluation protocol is the Grade, Roughness, Breathiness, Aesthenia, Strain (GRBAS) perceptual rating protocol [6]. It has been demonstrated that, on the basis of low intra-rater and inter-rater variances, the GRBAS protocol seems to be the most reliable and relevant perceptual voice quality evaluation [1].

On the other hand, the complexity analysis of pathological voices seeks to quantify the effects of nonlinear phenomena involved in the voice production process, due to changes in the dynamic properties of the vocal cords and laryngeal tissues because of the presence of pathology. This kind of analysis has demonstrated to provide more stable results than conventional acoustical analysis when the voice signals do not present a quasiperiodic structure [7]. Additionally, the information obtained using complexity analysis has demonstrated to be relevant for the evaluation of different types of laryngeal pathologies [7], and also complementary to the one obtained using conventional methods of characterization (such as noise measures and cepstral coefficients) for the automatic detection of pathological voices [8].

In this sense, this work explores the discrimination capabilities of nine complexity measures for the perceptual evaluation of pathological voices according to the GRBAS protocol and their use in an automatic system for the assessment of voice quality. Since each scale of the GRBAS protocol can take one of four different values (classes), rating a voice according to it, can be seen as a multiclass problem Therefore, the classification is carried out using three different multiclass strategies based on binary Gaussian Mixture Models (GMM) classifiers: One *vs* All, All *vs* All and Hierarchical (like tree) classification.

The results are shown in terms of efficiency and statistical agreement index. The subset of complexity

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

measures along with the classification strategy which provided the best result for each scale of the GRBAS protocol are also reported.

II. Methods

In the first stage of the process, the speech signal is framed and windowed in order to perform a short-time analysis. This approach is well established in speech processing tasks, including speech recognition, or speaker identification and verification. Nevertheless, the nonlinear analysis of speech signals on a frame basis is a recent approach [8]. This analysis is supported by the fact that changes in the dynamics of a pathological voice can be presented during long periods of time or suddenly. Slow changes in the speech signal are related to the biological processes in which the properties of the tissues evolve. On the opposite side, sudden changes can be explained by the presence of extra masses or changes in the biomechanical properties of the tissues of the vocal folds, modifying the dynamic behavior during the voice production process, and producing abrupt variations in the vibration regime of the vocal fold that can be understood as bifurcations [9]. These phenomena can be better detected and characterized using a short-time scheme. The analysis was carried out using frames of 55 ms with a 50% frame shift according to previous results [8], in which the frame length was selected taking into account criteria related to the minimum signal length that must be used for a good estimation of nonlinear features and also a minimum number of pitch periods for a good characterization of the signal stability. In the following section each of the complexity measures used in the characterization stage will be exposed.

A. Parameterization

First of all, a complexity analysis of biomedical signals requires a previous reconstruction of the state space of the underlying system to be characterized. Such reconstruction is carried out using a mathematical procedure called *embedding*, which typically is based on the time-delay embedding theorem [10]. The embedding theorem establishes that, when there is only a single sampled quantity from a dynamical system, it is possible to reconstruct a state space that is equivalent to the original (but unknown) state space composed of all the dynamical variables. The points in the state-space form trajectories, and the set of trajectories from a time series is known as *attractor*.

From each speech frame an attractor is reconstructed and subsequently a set of 9 complexity measures are estimated.

Largest Lyapunov Exponent (LLE): LLE is a measure of the separation rate of infinitesimally close trajectories of the attractor [10]. In other words, LLE measures the sensibility to the initial conditions of the underlying system, since one of the main characteristics of nonlinear systems is the possibility that two trajectories in the state space begin very close and diverge through time, which is a consequence of the unpredictability and inherent instability of the solutions in the state space. Theoretically, a positive value of LLE means an exponential divergence of nearby trajectories and consequently a more complex dynamic behavior in the attractor.

Correlation dimension (CD): CD is a measure of the dimensionality of the space occupied by a set of random points or its geometry. Moreover, it characterizes the scaling properties of a distribution of points in an *m*-dimensional space (being *m* the dimension of the embedded attractor). The CD is the fractal dimension that has received more attention in the literature. This is mainly because its estimation is easier than others. Besides, it provides a good measure of the complexity of the dynamics, i.e. it measures the number of active degrees of freedom.

Approximate Entropy (A_E): In the field of nonlinear dynamics, complexity measures often quantify statistically the evolution of the trajectory in the embedded phase space. However, if a signal is considered as the output of a dynamical system in a specific time period, it is regarded as a source of information about the underlying dynamics; therefore, the amount of information about the state of the system that can be obtained from the signal can also be considered as a kind of complexity. The fundamental idea to measure the "amount of information" comes from the information theory, and is termed *Entropy*. Entropy is a measure of the uncertainty of a random variable [8]. The most employed measure in this context is $A_{E,}$, which is a measure of the average conditional information generated by diverging points of the trajectory [8]. The advantage of using entropy based measures is that they measure the complexity of the signal without making assumptions about the nature of the process (deterministic or stochastic), whilst conventional nonlinear statistics such as LLE and CD assume that this nature is entirely [11], which cannot be asserted for voice signals.

There are several modifications of A_E published in the literature. Among them the most important is the *Sample Entropy (S_E)*, developed with the aim of obtaining a more independent measure than A_E with respect to the signal length.

Recurrence and fractal scaling analysis: Considering that there is a combination of both deterministic and stochastic components in the voice signal during phonation [11], the deterministic component can be

characterized by a measure called *Recurrence period density entropy (RPDE*) and the stochastic component by means of a *Detrendend fluctuation analysis (DFA)*. RPDE quantifies any ambiguity that might exist in the fundamental frequency; the level of ambiguity is often an indicative of vocal dysfunction [11]. On the other hand, DFA characterizes the changing details of aeroacoustic breath noise in the voice and therefore it is sensitive to similar features in voice as *Noise to Harmonic Ratio (NHR)*, but instead of NHR, DFA does not depend on a previous pitch estimation which is a difficult task for pathologic signals.

Hidden Markov entropy measurements: Most of the complexity measures used in the state of the art to characterize pathological voices, are based on multiple comparisons of the points in the attractor to establish the neighborhood of each point according to a particular distance measure. From such comparisons, the diverging points of the attractor are determined. The neighborhood of a particular vector in the state space is then understood as a region of the space in which the distance between that vector and the others is lower than a certain value (r). However, the temporal information of the points in the attractor is not taken into account. Since the points in the attractor should follow an ordered path –at least with normal stable voices–, the *Hidden Markov entropy measurements* were formulated to quantify the amount of information about the state of the system, taking into account the dynamic information of the points in the attractor [8]. The dynamic of the points in the attractor is modeled as a *hidden Markov process (HMP)* throughout a *discrete hidden Markov model (DHMM)*, which can also be seen as an estimation of the probability density function of the process; from this model three different entropy measures are estimated: the entropy of the Markov chain (H_{MC}), and two empirical estimations of the DHMM entropy: Shannon entropy (H_{ES}) and Renyi entropy (H_{ER}).

All the complexity measures described in this section have already been used for the characterization of voice diseases and also for the automatic detection of pathological speech signals [7,8,11], showing relevant results.

B. Classification

As previously commented, each scale of the GRBAS protocol can take one of four different values (classes), therefore the classification of a voice according to such protocol can be seen as a multiclass classification problem. In this sense, the classification in this work is performed using three different multiclass strategies based on binary classifiers, namely One-*vs* All, All-*vs*-All and Hierarchical, all of them employing GMM as the core of the pattern classifier stage.

One vs all: In this approach, a binary classifier discriminates between a given class and the other $nc-1$ classes. For this approach, the number of binary classifiers required is $N=nc$, where the k-th classifier is trained with positive examples belonging to class k and negative examples belonging to the other $nc-1$ classes. When testing an unknown pattern, the classifier that provides the maximum output is considered the winner, and the label of this class is assigned to that pattern.

All vs all: In this approach, a binary classifier is built to discriminate between every possible pair of classes, while discarding the rest of the classes. This requires $N=nc\cdot(nc-1)/2$ binary classifiers. When testing a new example, a voting is performed among the classifiers and the class with the maximum number of votes wins.

Hierarchical: Another way to address the multiclass classification problem is to perform a hierarchical division of the output space, i.e. arranging the classes like a tree. The tree is created in such a way that the classes at each parent node are divided into a number of clusters, one for each child node. The process continues until the leaf nodes contain only a single class. At each node of the tree, a simple classifier, usually a binary classifier, makes the discrimination between the different child class clusters. Following a path from the root node to a leaf node leads to a classification of a new pattern. This method uses $N=n_c-1$ binary classifiers for an n_c-class problem

C. Experimental Setup

Testing was carried out using a subset of the database developed by The Massachusetts Eye and Ear Infirmary Voice & Speech Laboratory. All available 226 voices (173 pathological and 53 normal) were presented to an experienced voice therapist in a randomized order and without providing any information about the diagnosis. For each speaker, both recordings (sustained vowel and running text) were made available to him and he was asked to provide a perceptual rating for each speaker according to the GRBAS protocol. The validation was performed using a leave-one-out crossvalidation strategy due to the small number of voice recorders belonging to some of the classes.

.

III. RESULTS

Table I shows the set of complexity measures which perform better for each scale of the GRBAS protocol. From these results it is possible to know which characteristics provide the largest contribution to the automatic evaluation of each scale of the GRBAS protocol. The best sets of features were determined based on a brute-force search and discriminative criteria. Table

I also shows the efficiency achieved for every scale. It is also worth to note that, in most of the cases, the best performance was obtained by using a "One-*vs*-All" multi-class classification strategy.

Table I. Best sets of complexity measures and efficiency obtained for each scale of the GRBAS protocol.

Scale	*Set of features*	*Efficiency [%]*
G	*CD, A_F, S_F, DFA, H_{MC}*	56.44
R	*LLE, CD, A_F, S_E*	55.11
B	*LLE, S_E, DFA, H_{ES}, H_{ER}*	57.18
A	*LLE*	66.67
S	*H_{MC}*	46.67

Table II shows estimations of the statistical agreement index, Kappa, for each scale of the GRBAS protocol. The Kappa index measures the agreement between the classification provided by the system and the evaluation supplied by the specialist who labeled the database. For the sake of comparison, table II also shows the agreement obtained between to experienced medical specialist for the assessment of pathological voices according to the GRBAS protocol reported in [1]. The results show that although the agreement obtained by the system is still lower than the one obtained by two specialists, the results are comparable and therefore the information obtained from the complexity analysis and the classification methodology employed in this work can be useful for improving the automatic assessment of voice quality according to the GRBAS protocol.

Table II. Kappa indexes obtained by the system for each scale of the GRBAS protocol.

Scale	*Kappa*	*Kappa in [1]*
G	0.40	0.51
R	0.40	0.46
B	0.37	0.43
A	0.32	0.41
S	0.24	0.34

IV. Discussion and Conclusions

The analysis of agreement between the automatic system and the rater who labeled the database showed that the performance of the system is a bit lower compared to the agreement obtained by two experienced specialist reported by [1]. Nevertheless, the results show that complexity analysis provides relevant information for this task. It is worthy to note, that the nonlinear analysis of speech signals is proposed as a complement of the analysis based on classical acoustic parameters. Therefore, it is very likely that, similar results to the obtained in [8] for the detection of pathological voices, can be reached by means of the combination of conventional and nonlinear analysis, improving the automatic rate of voices according to perceptual criteria. The multiclass classification strategy showed interesting results, however, it remains open the problem of dealing with a small number of samples in some classes.

Acknowledgments

This research work has been financed by the Universidad Antonio Nariño, Colombia, under grant 2010238 - PI/UAN-2011-476bit and Spanish government through the project grant TEC2009-14123-C04-02.

References

[1] P. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Wolsard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," *Rev. Laryngol. Otol. Rhinol*, vol. 117, no. 3, pp. 219–224, 1996.

[2] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[3] T. Butha, L. Patrick, and J. D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements," *Journal of Voice*, vol. 18, no. 3, pp. 299–304, 2004.

[4] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[5] R. Ritchings, M. McGillion, and C. Moore, "Pathological voice quality assessment using artificial neaural networks," *Medical Engineering &Physics*, vol. 24, no. 8, pp. 561–564, 2002.

[6] M. Hirano, *Clinical Examination of Voice*. New York, USA: Springer-Verlag, 1981.

[7] J.J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement", *Journal of Voice*, vol. 20, no. 1, pp 2–17, 2006.

[8] J. Arias-Londoño, J. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters and mel-cepstral coefficients," *IEEE Trans. on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.

[9] Titze, I. R., *The Myoelastic Aerodynamic Theory of Phonation*. The National Center for Voice and Speech, Iowa, IA, USA, 2006.

[10] Kantz, H. and Schreiber, T., *Nonlinear time series analysis*, 2nd ed, Cambridge University Press, UK, 2004.

[11] Little, M.A., McSharry, P. E., Roberts, S. J., Costello, D. A., and Moroz, I. M. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection". *Biomedical Engineering Online*, vol. 6, no. 23, 2007.

ESTIMATION OF HARMONIC AND NOISE COMPONENTS OF THE GLOTTAL EXCITATION

R. Sousa[1], A. Ferreira[1] and P. Alku [2]
[1]*University of Porto – School of Engineering, Porto, Portugal*
[2] *Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*

Abstract: **This paper describes an algorithm which enables harmonic and noise splitting of the glottal excitation of voiced speech. The algorithm utilizes a straightforward harmonic and noise splitter which is utilized prior to glottal inverse filtering. The results show improved estimates of the glottal excitation in comparison to a known inverse filtering method.**
Keywords: **Voice quality, voice diagnosis, glottal inverse filtering**

I. INTRODUCTION

Since the glottal volume velocity waveform serves as the source of (voiced) speech, it has an essential role in the production of several acoustical phenomena such as the regulation of vocal intensity [1], voice quality [2], the production of different vocal emotions [3] and voice pathologies detection related to vocal fold changes [4]. Therefore, accurate analysis and parameterization of the glottal pulseform is beneficial in several areas of speech science including both healthy and disordered voices. In this paper, two techniques are combined to yield an algorithm that estimates the harmonic and noise components of the glottal pulse. These techniques decompose the signal into a harmonic and noise component and gives rise to better glottal pulse estimations. This new algorithm was tested with synthetic and natural voices in order to characterize the algorithm behavior against an acoustic diversity.

II. METHODS

A. Algorithm overview

The main goal of the study is to develop an algorithm that splits the waveform of the estimated glottal airflow into a harmonic and a noise component. The block diagram of the method is shown in Fig. 1.

First (block 1), the speech pressure signal is divided into a harmonic and a noise component [5]. Secondly (block 2), the obtained harmonic component of the speech signal, denoted by h(n) in Fig. 1, is used as an input to glottal inverse filtering which yields an estimate of the vocal tract inverse filter (an FIR filter), denoted by V(z) in Fig. 1. Inverse filtering is computed with a previously developed automatic algorithm, Iterative Adaptive Inverse Filtering (IAIF) [6]. Thirdly, this FIR

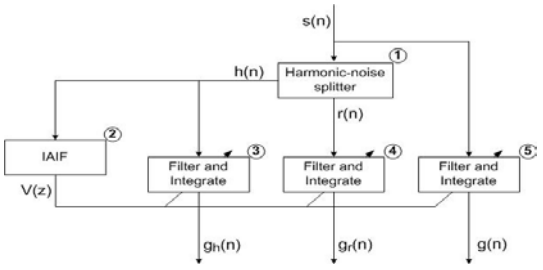


Fig. 1: Main block diagram of glottal harmonic-noise splitter. Signals s(n), h(n) and r(n) denote, respectively, the speech signal and its harmonic and noise components. Signals g(n), $g_h(n)$ and $g_r(n)$ denote, respectively, the glottal excitation, and its harmonic and noise components. V(z) denotes the vocal tract transfer function. IAIF denotes the glottal inverse filtering algorithm [6].

filter is used in order to cancel the effects of the vocal tract from three signals: both from the harmonic and noise components obtained from the harmonic-noise splitter, and from the original speech pressure waveform. By further canceling the lip radiation effect using an integrator whose transfer function is simply given by $H(z)=1/(1-0.99z^{-1})$, three glottal signals are obtained: the glottal pulse harmonic component, the glottal pulse noise component, and the glottal pulse, which are denoted in Fig. 1 by $g_h(n)$, $g_r(n)$, and g(n), respectively. Equations (1) to (4) express the resulting signals in Fig. 1.

$$s(n) = h(n) + r(n) \quad (1)$$

$$g(n) = v(n) * \ell(n) * [h(n) + r(n)] \quad (2)$$

$$g(n) = v(n) * \ell(n) * h(n) + v(n) * \ell(n) * r(n) \quad (3)$$

$$g(n) = g_h(n) + g_r(n) \quad (4)$$

The parameters *v(n)* and $\ell(n)$ denote the impulse response of the inverse model of the vocal tract and lip radiation effect, respectively. Equation (1) represents the harmonic-noise model, which serves as the basis for the harmonic-noise splitter. Inverse filtering is represented by equation (2). Equations (3) and (4) show that the glottal excitation consists of harmonic and noise components.

The harmonic-noise splitter is based on a model of the harmonic structure of speech, which is parameterized in frequency, magnitude and phase [5]. The block diagram of the harmonic-noise splitter is depicted in Fig. 2.

In the first stage (block 1), the time domain input

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

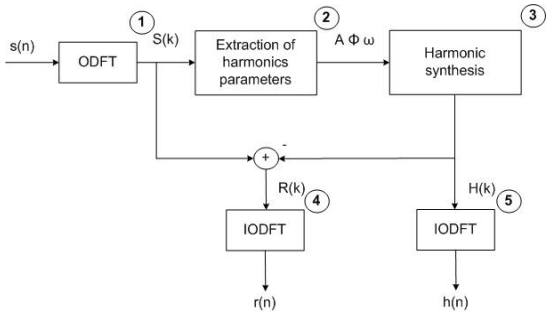


Fig. 2: Block diagram of the harmonic-noise splitter.

signal is transformed into the frequency domain using an Odd-Discrete Fourier Transform (ODFT) [7]. ODFT is obtained by shifting the frequency index of the Discrete Fourier Transform (DFT) by half a bin:

$$X_o(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}(k+\frac{1}{2})n}, \quad k = 0,1,...,N-1 \qquad (5)$$

where the time-domain input signal is denoted by x(n) and the frame length is N. If x(n) is real, this frequency shift makes the DFT samples above π a perfect mirror (in the complex conjugate sense) of the DFT samples below π. A peak picking algorithm is used to estimate the harmonics of the ODFT amplitude spectrum. Next, the frequency, magnitude and phase of each harmonic are extracted (block 2) [7]. These parameters are then used to synthesize the spectrum of the harmonic structure of the input signal s(n) (block 3). The spectrum of each individual sinusoid is synthesized using the parameters extracted from that harmonic.

The synthesized harmonic structure is subtracted from the signal s(n) and the result is regarded as the noise component. The spectra of both components are inverse transformed in order to get time-domain representations for the components (blocks 4 and 5).

B. Performance assessment

Experiments were conducted by using both synthetic and natural vowels. The estimated glottal excitation waveforms were parameterized with two known parameters: the Normalized Amplitude Quotient (NAQ) and the difference (in dB) between the amplitudes the first and second harmonic (DH12). The NAQ parameter is a time-based parameter that is extracted for each glottal pulse and it measures the pressedness of phonation from the ratio of the peak-to-peak flow and the negative peak amplitude of the flow derivative [8]. The DH12 parameter is a frequency domain quantity and it measures the decay of the voice source spectrum [9]. Both parameters are independent of time and amplitude shifts. The relative error was used for NAQ since this parameter is a time-domain quantity that is typically measured on the linear scale and the absolute error was used for DH12 because this parameter is typically expressed in the dB scale.

A synthesizer based on the source-filter and harmonic-noise models was used to generate a set of test vowels. The source generation was based on Liljencrants-Fant (LF) model [10]. The fundamental frequency F0 was varied from 100 Hz up to 400Hz with an increment of 10 Hz, in order to mimic both male and female speech. For each pitch, several vowel instances were generated by varying HNR from 9 dB up to 21 dB with an increment of 1 dB. The HNR is acquired as:

$$HNR = 10 \times \log_{10}\left(\frac{E_h}{E_r}\right) \qquad (6)$$

E_h and E_r denote, respectively, the energy of the harmonic component and the noise component of synthetic speech. The values of the LF model were selected according to Gobl [11] in order to involve three different phonation types (breathy, normal and pressed). The vocal tract filter was adjusted to synthesize the vowel [a] (F1= 664 Hz, F2=1027 Hz, F3=2612 Hz). All the data were generated using the sampling frequency of 22.05 kHz.

In the second experiment, a database that included 39 sustained waveforms of the vowel [a] uttered by 13 subjects (7 males, 6 females) using breathy, normal and pressed phonation was used. The data were sampled with 22.050 kHz and a resolution of 16 bits. From these signals, the most stable segments with duration of 200 ms were selected for the voice source analysis.

III. RESULTS

A. Experiments with synthetic voices

This section presents the results that were obtained for synthetic voices when the glottal source was estimated with IAIF and the proposed method. The NAQ error and DH12 error were determined separately for each phonation type. In order to compress the results, a set of ranges were defined for F0 and HNR and the individual values obtained inside these ranges were pooled together. For F0, the following three ranges were used: 100-200 Hz, 210-300 Hz, and 310-400 Hz. The first two ranges correspond to typical pitch used by males and females, respectively. The third range represents F0 values typical in voices produced by children. For HNR, the following three categories were used: 9-15 dB, 16-21 dB, and 22-27 dB. The first of these is typical for pathological voices while the second is characteristic to normal speech [12]. The last HNR range is related to voices which are highly periodic with a small amount of noise, such as the singing voice [13]. For each phonation type, the results are organized in tables that show the performance of NAQ or DH12 for the selected F0 and HNR ranges.

Tables 1 and 2 show that the proposed algorithm yields smaller DH12 errors for all the F0 and HNR combinations analysed from pressed vowels. The mean NAQ error was smaller with the proposed method also for all the F0 and HNR combinations except for three

cases (F0 ranges 210-300 Hz and 310-400 Hz combined with HNR range of 16-21 dB; F0 range 310-400 Hz combined with HNR range 22-27 dB).

Table 1: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of pressed synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	27,8	14,8	22,6	13,0	11,2	15,5
210-300	52,8	27,5	75,6	21,2	38,4	60,4
310-400	64,7	68,9	131,3	55,9	101,1	151,0

Table 2: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of pressed synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	4,6	1,4	0,8	1,0	0,5	0,4
210-300	14,3	3,6	4,0	4,7	2,4	2,0
310-400	15,0	15,1	7,8	12,3	4,7	5,9

Tables 3 and 4 indicate that the proposed method yielded smaller errors for all the F0 and HNR ranges in the NAQ measurements in modal phonation.

Table 3: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of modal synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	38,2	21,3	9,3	14,2	8,0	4,7
210-300	68,9	38,2	16,7	24,4	11,4	10,8
310-400	68,5	54,5	36,5	38,3	24,0	28,0

Table 4: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of modal synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	7,2	0,9	0,8	1,6	1,4	0,7
210-300	15,7	6,4	3,8	5,4	1,0	1,9
310-400	9,4	16,3	11,9	16,9	4,0	2,9

For the DH12 error, the proposed method yielded larger distortion than IAIF only in two cases (F0 range of 100-200 Hz combined with the HNR range of 16-21 dB; F0 range of 310-400 Hz combined with HNR range of 9-15 dB).

Tables 5 and 6 show results from breathy voices that are in line with those observed for modal phonation: the mean NAQ error is smaller for the proposed method for all the F0 and HNR categories analysed and the mean DH12 error was also smaller with the proposed algorithm in comparison to IAIF for all the F0 and HNR combinations except for few cases (F0 range of 100-200 Hz combined with the HNR ranges of 16-21 dB and 22-27 dB; F0 range of 210-300 Hz combined with HNR range of 22-27 dB).

Table 5: NAQ mean relative error (in percentage) for IAIF and the proposed method in the analysis of breathy synthetic voices.

F0 (Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	56,9	37,0	16,5	25,8	11,6	12,0
210-300	77,9	68,2	23,9	46,9	17,9	13,3
310-400	83,8	80,7	45,8	54,4	31,6	18,9

Table 6: DH12 mean absolute error (in dB) for IAIF and the proposed method in the analysis of breathy synthetic voices.

F0(Hz)	IAIF HNR (dB)			Prop. Meth. HNR (dB)		
	9-15	16-21	22-27	9-15	16-21	22-27
100-200	9,8	4,6	2,4	5,3	5,1	4,3
210-300	32,8	24,3	4,5	15,7	6,7	5,5
310-400	21,0	28,2	13,3	20,8	9,1	5,7

In summary, the results obtained for the synthetic vowels show that the proposed method yields smaller mean NAQ and DH12 errors for the majority of the sounds analyzed. In particular, we highlight that the proposed method yields improved estimation accuracy in conditions with large amount of noise and for high-pitch voices. This accuracy improvement depends on the phonation type being more pronounced for modal voices.

B. Experiments with natural voices

Results computed from natural speech are shown in the form of time-domain waveforms by involving both the harmonic and the noise component yielded by the novel inverse filtering method.

Figures 3 and 4 show waveforms computed from utterances produced by a male and female speaker, respectively. From both of these figures one can observe that the harmonic component is smoother than the glottal excitation waveform. In addition, low frequency fluctuations are not present in the harmonic component and the noise component indicates amplitude perturbations at the instants of glottal closure.

IV. Discussion

Results obtained with synthetic voices show that the proposed method improves the estimation of the glottal waveform. The harmonic component given by the new algorithm is a more accurate estimate of the glottal source because the method is able to suppress the influence of noise which is always present in natural speech, particularly in pathological voices. The behavior of both algorithms was tested as a function of the noise level and fundamental frequency. The proposed method also

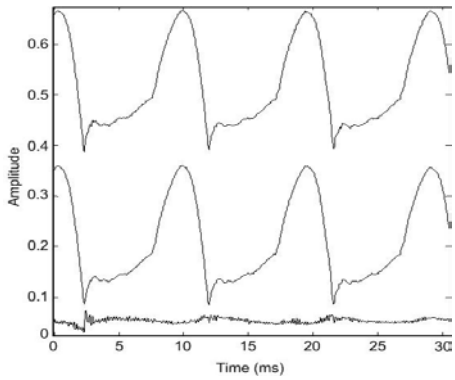


Fig. 3: Glottal excitation (top), its harmonic (middle) and noise (bottom) components estimated with the proposed method. A natural vowel [a] produced by a male speaker was used. The noise waveform is magnified 3 times for visual clarity.

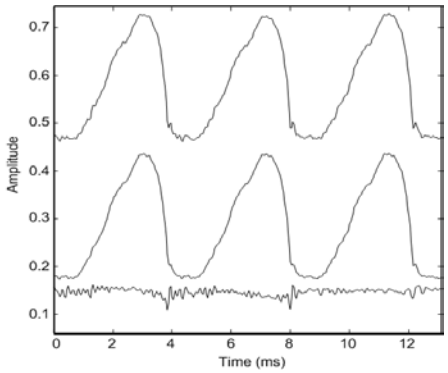


Fig. 4: Glottal excitation (top), its harmonic (middle) and noise (bottom) components estimated with the proposed method. A natural vowel [a] produced by a female speaker was used. The noise waveform is magnified 3 times for visual clarity.

enables joint estimation of the harmonic and noise components of the glottal waveform.

Drawbacks of the proposed method are due to the harmonic-noise splitter, which may pass noise to the harmonic component and itself is also sensitive to the noise level.

V. Conclusion

In this article, a method to estimate the glottal excitation based on a known automatic inverse filtering method, IAIF, and a harmonic-noise splitter was proposed. The new method was compared with IAIF in the estimation of the glottal excitation using experiments with both synthetic and natural vowels.

The proposed method enables joint estimation of the harmonic and noise components of the glottal waveform. These components may be used in the evaluation of pathological voices since the separation enables characterizing the vocal folds dynamics as a function of noise produced in the speech production process. In addition, the noise component estimated by the proposed method can be used in speech technology in order to improve the naturalness of synthetic speech.

Acknowledgements

This work has been developed in the context of a doctoral program supported by the Portuguese Foundation for Science and Technology under grant SFRH/BD/24811/2005 and project PTDC/SAU-BEB/104995/2008.

References

[1] F. Hodge, R. Colton and R. Kelley, "Vocal intensity characteristics in normal and elderly speakers," *J. Voice*, vol.15, no.4, pp. 503–511, 2001.

[2] C. Gobl and A. Ni Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, pp. 189–212, 2003.

[3] K. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Commun.*, vol. 40, pp. 227–256, 2003.

[4] P. Vilda, R. Baillo, V. Biarge, V. Luis, A. Marquina, L. Fernandez, R. Olalla and J. Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Commun.*, vol. 51, pp.759–781, 2009.

[5] R. Sousa, "A new accurate method of harmonic-to-noise ratio extraction," *Proc. of the Inter. Conf. on Bio-inspired Systems and Signal Proc.*, pp. 351-356, 2009.

[6] P. Alku, "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109-118, 1992.

[7] A. Ferreira, "Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids," *In: IEEE Workshop on App. of Signal Proc. to Audio and Acous.,* pp. 47–50, 2001.

[8] P. Alku, T. Bäckström and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, pp. 701-710, 2002.

[9] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Am.*,vol. 91, no. 5, pp. 2936–2946, 1992.

[10] G. Fant, J. Liljencrants, Q. Lin, "A four parameter model of glottal flow," *STL-QPSR*, vol. 1, pp.1-13, 1985.

[11] C. Gobl, "A preliminary study of acoustic voice quality correlates," *STL-QPSR*, pp. 9-21, 1989.

[12] J. Llorente, P. Vilda, F. Roldan, M. Velasco and R. Fraile, "Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness," *J. Voice*, vol.15, no. 7, pp. 1-11, 2009.

[13] J. Selby, H. Gilberta and J. Lerman, "Perceptual and acoustic evaluation of individuals with laryngopharyngeal reflux pre- and post-treatment," *J. Voice*, vol. 17, no. 4, pp. 557–570, 2003.

MULTISCALE PRODUCT CORRELATION FOR THE OPEN QUOTIENT ESTIMATION FROM THE NOISY SPEECH SIGNAL

Wafa SAIDI[1], Aïcha BOUZID[2] and Noureddine ELLOUZE[3]

[1, 2, 3] LSTS, ENIT-Tunis Le Belvédère B.P 37, 1002 Tunis, Tunisie +216 71 874 700, +216 71 872 729

Abstract: **In this paper, an improved Multiscale Product-based Method is evaluated for the open quotient (OQ) estimation from the noisy speech signal. The method consists of making the multi-scale wavelet transforms coefficients product at three scales. Our proposed approach is based upon correlation functions computed on the negative and the positive parts of the speech Multiscale Product (MP). It operates without determining the glottal opening and closing instants. Over each frame, the pitch period is given by the first maximum of the MP negative part autocorrelation and the open phase is given by the first maximum of the inter-correlation between the negative and positive parts. OQ is the ratio of the open phase and the pitch period. Tested on the Keele University database, our new approach proves to be robust to noise degradation.**

Key words: **open quotient, speech, multi-scale product, correlation, noise.**

I. INTRODUCTION

According to the acoustic theory of the speech production, the acoustic source signal produced by the vibrating vocal folds is filtered by the vocal tract to produce the speech output signal [1]. For voiced speech, the glottal vibration is periodic, with the folds opening and closing repeatedly in a regular manner. Thus, one period of the voice source signal includes open phase and closed phase. During the closed phase, the vocal folds are in full contact and there is no air flow passing through the glottis. The open phase is itself divided into an opening phase during which the vocal cords begin to separate gradually and a closing phase during which the separated folds start to be in close. Therefore, during the open phase the air passes through the glottis and the vocal cords are totally or partially detached.

The instant of vocal folds full contact is called the glottal closure instant (GCI) and one of vocal folds complete separation is the glottal opening instant (GOI). GCI and GOI are events of great interest for the glottis excitation. The open quotient is another interesting parameter characterising the source signal. It is defined as the ratio between the open phase and the cycle period.

Inverse filtering (IF) is a common and useful technique for voice source analysis. The principle of the IF is to cancel the vocal tract effect from a recorded speech signal to acquire a glottal flow [2].

Direct measurement of the glottis parameters from the radiated speech signal is still a challenging problem in speech analysis and synthesis domains. Though, numerous parameterisation approaches have been suggested. Time-based methods consist of detecting significant events such as glottal opening and closing instants to compute the glottis parameters [3], [4]. Frequency-based methods use the properties of the flow magnitude spectrum such as the level difference of the harmonics [5], [6]. In [7], Hanson uses the difference between the magnitudes of the first two spectral harmonics (H1 –H2) as an indication of the open quotient.

The electroglottographic recordings are used by many researchers to extract the glottal source features. Recently, Henrich et al. have proposed a correlation-based method called DECOM [8]. Her algorithm uses the correlation of the DEGG signal to estimate the fundamental frequency (F0) and the open quotient (OQ).

In this study, we focus on applying the Henrich correlation algorithm on the speech MP to estimate the open quotient from a noisy speech. The idea is born from the fact that the speech MP is strongly close to the EGG signal.

This paper is organised as follows. Section 2 reviews the principle of the multi-scale product analysis. Section 3 describes the Correlation Multiscale Product-based method for measuring the open quotient from the speech signal. In section 4, we evaluate the performance of our approach on clean and noisy speech data. Section 5 concludes this work.

II. MULTISCALE PRODUCTS FOR SPEECH ANALYSIS

Wavelet transform is a multiscale analysis widely used in image and signal processing. Due to the efficient time-frequency localisation and the multiresolution characteristics, the wavelet transforms are quite suitable for processing signals of transient and non-stationary nature. In [9], Mallat has shown that

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

multiscale edge detection is equivalent to find the local maximum of its wavelet representation. Glottal closure and opening instants are such events characterising the speech signal. The peak displaying the discontinuity in the wavelet transform is often damaged by noise when the scale is so fine or smoothed when the scale is large.

To improve edge detection using wavelet analysis, the multiscale product method is proposed. The latter consists of making the product of the wavelet transform coefficients of the acoustic signal over three scales. It enhances the peak amplitude of the modulus maxima line and eliminates spurious peaks due to the vocal tract effect.

The product of the wavelet transform of a function $f(n)$ at scales is:

$$p(n) = \prod_j W_{Sj} f(n) \qquad (1)$$

Were $W_{Sj} f(n)$ represents the wavelet transform of the function $f(n)$ at scale sj.

The product $p(n)$ shows peaks at signal edges, and has relatively small values elsewhere. An odd number of terms in $p(n)$ preserves the edge sign.

The MPM was first related to the edge detection problem in image processing [10]. Besides, the MPM is proposed by Bouzid and Ellouze to extract crucial information concerning the vocal source from both the speech and the electroglottographic signal (EGG) such as glottal opening and closure instants, the fundamental frequency, the open quotient and the voicing decision [11], [12].

III. MULTISCALE PRODUCT CORRELATION-BASED METHOD FOR OPEN QUOTIENT MEASUREMENT

As illustrated in Fig. 2, our proposed approach for the open quotient estimation from the speech signal operates following three stages. The first stage consists of computing the MP of a voiced speech signal and then dividing it into frames of a fixed length. The second stage consists of separating the speech MP into two parts: a negative part MPc which contains information concerning glottal closure peaks, and a positive part MPo which contains information about glottal opening peaks. The MPc signal is derived from the original signal by replacing any positive value by zero. In the same way, the MPo signal is derived from the original signal by replacing any negative value by zero.

The third stage concerns the calculation of the inter-correlation function between the positive and negative parts (MPo and MPc) to estimate the open phase, and the autocorrelation function of the MPc to estimate the pitch period over each frame. The open phase and the pitch period are respectively given by the non null index matching with the first maximum of the intercorrelation and autocorrelation functions. The open quotient is then deduced by calculating the ratio between the open phase and the pitch period.

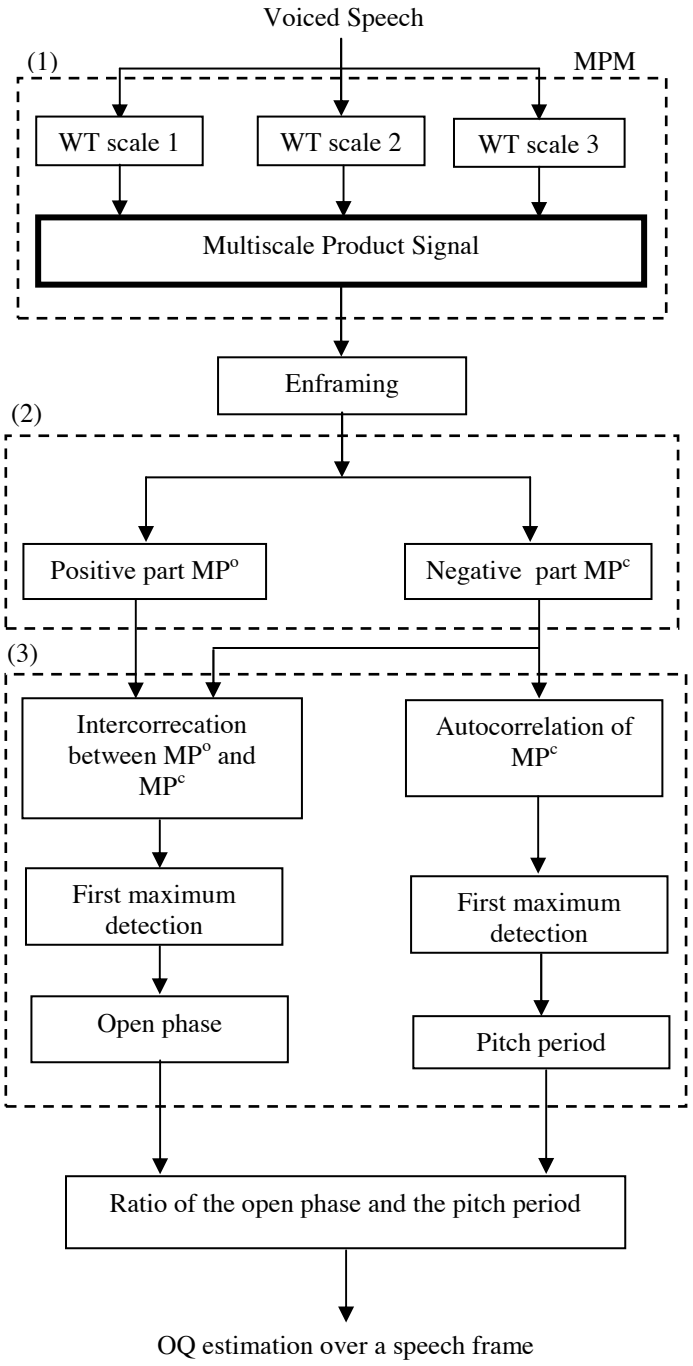


Fig. 1: *Overview of the proposed method.*

To compute the MP, we multiply the wavelet transforms of the speech signal at scales 2, 5/2 and 3 using the quadratic spline function.

To divide the MP signal into frames of a length N, we multiply it by a sliding rectangular window w[N]. The MP over a window of index i is given by the following equation:

$$MP_{wi}[k] = MP[k - iN]w[k] \qquad (2)$$

Where k is within [1, N] and i is the frame index.

The intercorrelation function between MPo and MPc over a frame i is calculated as follows:

$$R_o(k) = \sum_{l=1}^{N} MP_{wi}^{o}(l) MP_{wi}^{c}(k+l) \qquad (3)$$

As the same way, the autocorrelation function of MPc over a frame i is calculated as follows:

$$R_c(k) = \sum_{l=1}^{N} MP_{wi}^{c}(l) MP_{wi}^{c}(k+l) \qquad (4)$$

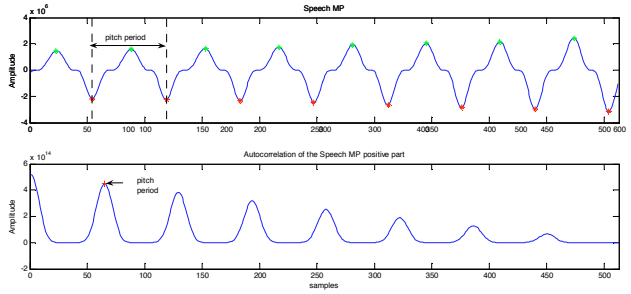


Fig. 2 : *Speech MP and the autocorrelation function of its positive part.*

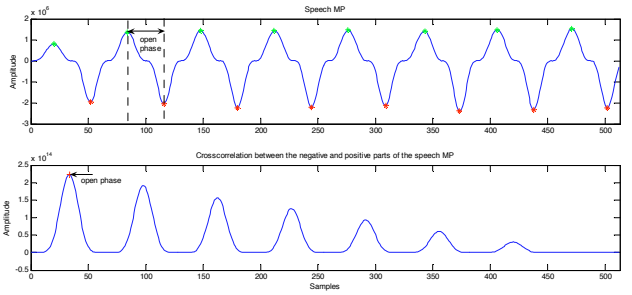


Fig. 3 : *Speech MP and the inter-correlation between its negative part and positive one*

The non null index matching with the first maximum of the MPc autocorrelation function shown in the second part of the Fig. 2 corresponds to the pitch period which is defined as the distance separating two consecutive GCIs.

Fig. 3 shows the speech MP followed by the inter-correlation calculated between its positive and negative parts. The non null index matching with the first maximum of the inter-correlation function corresponds to the time between an opening peak and the consecutive closing peak which is termed as the open phase.

IV. EVALUATION RESULTS

In this section, we evaluate the performance of our proposed method for OQ estimation using the Keele University database. This database includes the acoustic speech signals and laryngograph signals (single speaker recording). Five adult female speakers (fi) and five adult male speakers (mi) with $i \in \{1,...5\}$ are recorded in low ambient noise conditions using a sound-proof room. Each utterance consists of the same phonetically-balanced English text. In each case, the acoustic and laryngograph signals are time-synchronised and share the same sampling rate value of 20 kHz [13].

To evaluate the performance of our OQ estimator, we compute the standard deviation (σ) of the error measured as a difference between the OQ estimated from the speech and the EGG signals.

To study the noise effect on the accuracy of our open quotient estimator, we add noise to the original speech signal at various SNR levels. The noise is taken from the noisex-92 database [14]. Babble and vehicle noises are considered in this work.

Table 1 shows the performance of our approach for OQ estimation from the clean and noisy speech.

On clean speech approach estimates OQ with a standard deviation ranging from 0.03 for f2 to 0.08 form m5. It's a considerable accuracy for estimating open quotient from the speech signal. In fact, works developed in this field usually use the EGG recordings.

In the presence of noise at SNR levels ranging from 5dB to -5dB, we can notice that the noise has insignificant effect on the accuracy of the proposed approach. The majority of speakers save the same standard deviation value when adding noise. For others, the deviation increases finely when the SNR level reaches -5 dB.

V. CONCLUSION

In this paper, we have proposed an improved Multiscale Product-based method for estimating the open quotient from clean and noisy speech signal. The proposed method exploits the correlation of the speech multiscale product which reminds the derivative of the EGG signal shape representing the global source activity.

The OQ estimation is obtained by calculating the ratio of the open phase over the pitch period. The open phase is referred as the index non null of the first

maximum localised on the inter-correlation function between the positive and the negative parts of the speech MP. As the same way, the pitch period is the non null index matching with the first maximum of the speech MP correlation function.

Standard deviation between OQ estimated from the speech signal and OQ measured from the EGG signal is measured to evaluate our method. The evaluation is done on the Keele University database in a noisy environment. Noise is extracted from the noisex-92 database. The proposed approach is proved to be accurate and robust.

REFERENCES

[1] G. Fant, "Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations", *2nd ed. The Hague, Netherlands. Monton*, 1970.

[2] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow wavefoom during voicing", *J. Acoust Soc Am*, vol. 53, no. 6, pp. 1632-1645, 1973.

[3] J. Sundberg, I. Titze, and R. Scherer, "Phonatory control in mail singing: A study of the effects of the subglottal pressure, fundamental frequency, and mode of phonation on the voice source", *J. Voice 7,* 15-19, 1993.

[4] A. Bouzid, and N. Ellouze, "Open quotient measurements based on multiscale product of speech signal wavelet transform", *Research Letters in Signal Processing,* Vol. 2007, 2007.

[5] D. G. Childers, and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception", *J. Acoust Soc Am. 90,* 2394-2410, 1991.

[6] P. Howell, and M. Williams, "Acoustic analysis and perception of of vowels in children's and teenagers' stuttered speech", *J. Acoust Soc Am. 91,* 1697-1706, 1992.

[7] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", *J. Acoust. Soc. Am. 101,* 466-481, 1997.

[8] N. Henrich, C. d'Allessandro, M. Castellengo, and B. Doval, "On the use of the deravative of electroglottographic signals for characterization of nonpathological phonation", *J. Acoust. Soc. Amer.,* Vol. 115 (3), pp. 1321-1332, 2004.

[9] S. Mallat, and S. Zhong, "Characterization of signals from multiscale edges", *IEEE transactions on pattern analysis and machine intelligence,* Vol. 14, no. 7, pp. 710-732, 1992.

[10] A. Rosenfeld, "A Non Linear Edge Detection", *Proc.IEEE,* Vol.58, pp. 814-816, 1970.

[11] A. Bouzid, and N. Ellouze, "Voice source measurement based on multiscale analysis of electroglottographic signal", *10.1016/j.specom.2008.08.004, Speech Communication, Elsevier Publisher.*

[12] A. Bouzid, and N. Ellouze, "Electroglottographic measures based on GCI and GOI detection using multiscale product", *international journal of computers, communications and control*, Vol. III, no. 1, pp. 21-32, 2008.

[13] G.Meyer, F. Plante, and W. A. Ainsworth, " A pitch extraction reference database", *The 4th European conference on speech communication and technology,* p. 837-4, 1995.

[14] Noisex92. In: Signal Processing Information Base (SPIB). The Signal Processing Society and the National Science Foundation. 2007. http://spib.rice.edu/spib/select_noise.html. Accessed 2 March 2010.

Table 1 : Performance of Multiscale Product Correlation method for open quotient estimation using the Keele database on clean and noisy speech

Noise Type	SNR	Standard Deviation (σ)									
		Female speakers					Male speakers				
		F1	F2	F3	F4	F5	M1	M2	M3	M4	M5
Clean speech		0.06	0.06	0.05	0.03	0.04	0.07	0.04	0.04	0.05	0.08
Babble	5 db	0.07	0.07	0.06	0.03	0.05	0.07	0.04	0.04	0.05	0.08
	0 db	0.07	0.07	0.06	0.03	0.05	0.07	0.04	0.04	0.05	0.08
	-5 db	0.08	0.07	0.06	0.03	0.05	0.07	0.05	0.04	0.05	0.08
Vehicle	5 db	0.07	0.06	0.05	0.03	0.05	0.07	0.04	0.04	0.05	0.08
	0 db	0.07	0.06	0.05	0.03	0.05	0.07	0.04	0.04	0.05	0.08
	-5 db	0.07	0.06	0.05	0.04	0.05	0.08	0.04	0.04	0.05	0.09

Special Session:
Innovative ways for acoustic analysis of non-quasi-periodic voices

Chairperson and Introduction:
P. Dejonckere

SPECIAL SESSION

INNOVATIVE WAYS FOR ACOUSTIC ANALYSIS OF NON QUASI-PERIODIC VOICES

P. H. Dejonckere [1]

[1] Experimental ORL Cath. Univ. Leuven (B), Fed. Inst. Occup. Diseases Brussels (B), ORL-Phoniatrics Utrecht Univ. (Nl)

Objective measurement of the severity of dysphonia typically requires signal processing algorithms applied to acoustic recordings. Since Lieberman (1963) introduced the concept of perturbation analysis in the area of voice, the most dominant acoustic parameter in clinical practice is the classical jitter. However jitter measurements have some critical limitations. According to a widely accepted guideline, in sustained vowels of dysphonic voices, only perturbation measures less than about 5% (quasi-periodic voices) are reliable: this is related to period extraction methods.

This means that traditional acoustic analysis programs available for clinical use are not suited for quality assessment of strongly irregular voices, as substitution voices (voices not generated by two vocal folds, particularly after total/partial laryngectomy) or spasmodic dysphonias. The basic protocol for multidimensional voice assessment as recommended by the European Laryngological Society (Dejonckere et al., 2001) specifically mentions that it is not suitable for a few very special categories of voices, as substitution voices and spasmodic dysphonia. Nevertheless a valid quality evaluation is essential for substitution voices, as in laryngeal oncology there may be different therapeutical options comparable in survival rate for the same nature and stage of cancer. In such cases, functional outcomes (voice, respiration, swallowing) gain major significance.

The strong irregularity that characterizes the substitution voices is the major problem for usual acoustic analysis.

This special session deals for a part with successful improvements of the traditional approach of the cycle-to-cycle variability. A breakthrough was made possible by the development of a synthesizer of ‘ realistic ’ pathologic voices, that cannot be recognized by expert listeners from true patient ’ s voices, and where the jitter ‘ put in ’ is exactly known. This allows to check as well the ability of pattern recognition of the human visual system as the validity of new algorithms for period detection, in different conditions of noise. The practical result is that the traditional threshold limit value of 5 % for jitter measures may be transgressed under some conditions that will be discussed.

Furthermore, the question remains about the clinical value of perturbation measurements when analyzing running speech of patients with either substitution voices or spasmodic dysphonia. The same question is relevant for noise measurements. It still becomes clearer that the acoustic parameters that are in some way related to the selection of voiced/unvoiced parts of the signal are the most successful ones in discriminating either different types of substitution voices or therapeutical effect in spasmodic dysphonia.

Another problem is the presence of tremor in some pathological voices: this mainly concerns neurological voices, and particularly spasmodic dysphonia, a focal laryngeal dystonia.

The estimation of tremor attributes in a speech signal involves the accurate extraction of the signal that modulates the time-varying fundamental frequency. A new significant attribute of tremor is introduced. It derives from the time-varying characteristic of the modulation level, namely the deviation of the modulation level. The mean modulation level and its deviation are combined in a quality indicator trying to classify speakers according to the prevalence of tremor in their voice. This innovative approach can be tested on sustained vowels uttered by patients who suffer from spasmodic dysphonia before and after medical treatment.

REFERENCES

Dejonckere PH, Giordano A, Schoentgen J, Fraj S, Bocchi L, Manfredi C *“To what degree of voice perturbation are jitter measurements valid ? A novel approach with synthesized vowels and visuo-perceptual pattern recognition.”* Biomedical Signal Processing and Control, 2011 (in print).

Manfredi C, Giordano A, Schoentgen J, FrajS,Bocchi L, Dejonckere PH, “*Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools*” Biomedical Signal Processing and Control, 2011 (in print).

Dejonckere PH, Giordano A, Schoentgen J, Fraj S, Bocchi L, Manfredi *C “Validity of jitter measures in non*

quasi-periodic voices. Part I : Perceptual and computer performances in cycle pattern recognition." LogopedicsPhoniatricsVocology, 2011 (in print).

Manfredi C, Giordano A, Schoentgen J, Fraj S, Bocchi L, Dejonckere PH *"Reliability of voice analysis software tools for highly irregular signals Part II: the effect of noise"*LogopedicsPhoniatricsVocology, 2011 (in print).

Dejonckere P.H., Bradley P., Clemente P., Cornut G., Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M., Woisard V.,*"A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessments techniques"*, Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS), *Eur. Arch. Otorhinolaryngol.*258, pp.77-82, 2001.

Dejonckere PH, Neumann KJ, Moerman MBJ, Martens JP, *"Perceptual and acoustic assessment of adductor spasmodic dysphonia pre- and post-treatment with botulinum toxin"*, Proc. 3rd AVFA International Worshop, 18th-20th May 2009, Madrid (Spain).

ANALYSIS OF GLOTTAL CYCLE TREMOR AND JITTER BY EMPIRICAL MODE DECOMPOSITION

C. Mertens[1], F. Grenez[1], V.Boucher[2], J. Schoentgen[1,3]
[1]Laboratory of Images, Signals and Telecommunication Devices, Université Libre de Bruxelles, Brussels, Belgium
[2]Laboratoire de Sciences Phonétiques. Université de Montréal, Montréal, Canada
[3]National Fund for Scientific Research, Belgium

Abstract : **The presentation concerns a method for tracking cycle lengths in voiced speech and analysing vocal cycle lengh fluctuations. The tracking of cycle lengths is based on a dynamic programming algorithm, which does not request that the signal is locally periodic and the average period length is known a priori. The obtained cycle length time series is then decomposed into a sum of intrinsic mode functions by means of empirical mode decomposition. These mode functions are then assigned to three categories, which are cycle length jitter, cycle length tremor and trend owing to intonation and physiological tremor. The characteristics of slow and fast perturbations are reported.**

Keywords : **vocal frequency, vocal tremor, vocal jitter, speech salience analysis, empirical mode decomposition**

I. Introduction

In clinical applications of speech analysis, speech cycles are detected to measure their lengths and amplitudes with a view to investigating slow (vocal tremor) and fast (vocal jitter and shimmer) perturbations of vocal frequency and speech cycle amplitude. Often, such analyses are frame-based and the cycle detection rests on the recursive detection and storage of speech signal extrema that occur in the vicinity of the instants of glottal excitation. To enable this selection, one often assumes that voiced speech segments are pseudo-periodic so that the peaks can be selected one by one on the base of a prior estimation of the typical fundamental period. The assumption of quasi-equal spacing is, however, valid for modal voices only and not for pathological ones, which may be characterized by large cycle-to-cycle fluctuations in length or amplitude. Cycle insertion or omission errors may therefore occur, which bias the acoustic cues of cycle regularity.

Here, the speech cycle tracking does not rest on the assumptions that the speech signal is locally periodic and the average period length known a priori. We propose to track speech cycles via a multiscale analysis that assigns a salience to each signal peak. A signal peak is a signal sample whose left and right neighbours are smaller. The salience of a speech signal peak designates the time interval over which this peak is a maximum. The vocal cycle detection method relies on dynamic programming to extract a cycle sequence the length perturbations of which is minimal. The cost function involves the second order differences of successive speech cycle durations as well as the cycle peak saliences. The tracker does not rely on estimates of the typical cycle length, as opposed to existing proposals involving dynamic programming in the extraction of the vocal frequency or glottal cycle length [1]. The obtained cycle length time series is then decomposed into a sum of intrinsic mode functions by means of empirical mode decomposition [2]. These mode functions are then assigned to three categories, which are cycle length jitter, vocal tremor and trend owing to intonation and physiological tremor.

Section 2 explains speech sample saliences, the tracking of the speech cycles and the extraction of the empirical mode times series, as well as the corpora. Section 3 reports the results and discussion of the tracking of slow and fast perturbations in speakers affected by vocal fatigue as well as synthetic tremored sounds.

II. Method

A. Preprocessing

The speech signal is band-pass filtered by means of a finite response (FIR) filter with cut-off frequencies equal to $60Hz$ and $1000Hz$ to remove additive low frequency hum, additive noise owing to turbulence as well as high-frequency formants.

The speech signal is then upsampled to $F_s = 192kHz$ to enable the peak positions to be measured with a precision requested by the size of vocal jitter, which in modal voice is expected to be $< 1\%$ of the typical cycle length.

B. Speech sample salience analysis

The salience of a signal sample (which may be a signal peak or not) is defined as the length of the longest temporal interval over which the signal sample is a maximum. A property of the salience is that a sample with a large salience has not necessarily a large amplitude and vice versa. For instance, in voiced speech, speech cycles are often characterized by a prominent signal peak that is the effect of the glottal excitation. The salience of that peak is expected to be high irrespective of the evolving signal amplitude.

The speech sample saliences have been obtained via a multi-scale analysis based on a sliding analysis window of

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

length N. Fast windowed salience analysis involves speeding up the algorithm by computing left and right-hand saliences and computing saliences for a subset of samples only. Details are reported in [3].

The final speech sample saliences are comprised between 1 and $2N-1$. It is recommended to discard the $N-1$ first and last sample saliences, that are conditioned by the array boundaries. The sliding analysis window length must therefore be chosen so as to minimize the loss of information owing to the array boundaries and maximize the relevance of the window-determined saliences with regard to the goal of the multi-scale analysis.

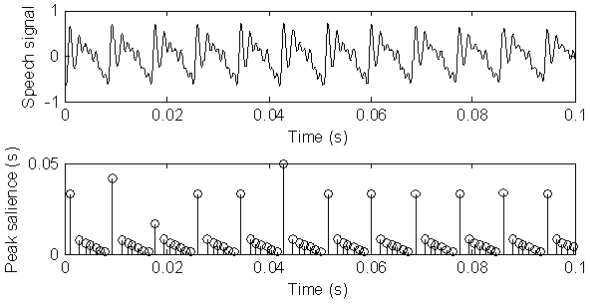


Fig. 1 : Signal peak saliences of sustained vowel fragment [a]

Fig. 1 illustrates the peak salience values obtained for a fragment of vowel [a]. One observes that the prominent signal peaks due to the glottal excitation have a higher salience value than other secondary peaks that are due to tract resonances.

C. Speech cycle tracking

For speech cycle tracking, no strong assumptions are made with regard to the regularity of the cycle lengths. One assumes that the vocal frequency is comprised between $60Hz$ and $400Hz$.

The first stage consists in ranking the signal peaks according to decreasing salience and keeping those peaks the salience values of which are greater than or equal to 150% the length of the shortest possible cycle. The initial number of peaks is therefore in excess of the number of expected cycles because a typical salience value of a speech cycle peak is equal to twice the cycle length.

The second stage consists in considering several candidate cycle length time series obtained by means of the retained peak distances and discovering via dynamic programming the length series that has the smallest overall cycle duration perturbation. The candidate cycle length series are built by taking into account several signal peak sub-sequences on the base of the local inter-peak durations and the peak salience values, assuming that prominent speech cycle peaks owing to the glottal excitation are characterized by large salience values. This second stage comprises an initialization, search and backtracking step. Details are reported in [4].

D. Vocal Jitter, vocal tremor and trend

The obtained vocal cycle length time series is then constant-step interpolated by :

1) reconstructing the temporal axis as the sum of the successive vocal cycle lengths,
2) interpolating the obtained series by means of cubic splines and
3) resampling to obtain a time series of lengths sampled at a constant sampling step.

The interpolated cycle length time series is decimated to a sampling frequency equal to 150% of the average vocal frequency. The decimated cycle length time serie is then decomposed iteratively into several intrinsic mode functions (IMF) by means of empirical mode decomposition [2], as follows :

$$x(n) = \sum_{i=1}^{M} IMF_i(n) + r(n) \quad (1)$$

where $IMF_i(n)$ is the i^{th} alternating function with respect to the zero local mean and $r(n)$ is a monotonic function, called residue. For each IMF_i, the spectrum and the abscissa f_G of the center of gravity of the spectrum are computed. According to the value of f_G, the IMF_i is then assigned to three categories : trend ($f_G < 3Hz$), vocal tremor ($3Hz < f_G < 15Hz$) and vocal jitter ($f_G > 15Hz$) and then added per category. Fig 2 reports the result of this assignation for the speech cycle time series of a French sustained vowel [a], for instance.

E. Vocal cues

One acoustic cue is the abscissa of the center of gravity of the low-frequency amplitude spectrum of the vocal tremor cycle length series, computed in the frequency interval [3-15Hz]. The boundary is fixed at $3Hz$ because cardiac beat, breathing and bloodflow are expected to influence strongly the spectrum below 3Hz. Two other cues are the standard deviation of the vocal tremor or vocal jitter cycle length time series divided by the average cycle length. These cues characterize the excursion of the cycle durations with respect to their average. The three cues are estimates of respectively the modulation frequency and modulation depth owing to tremor as well as jitter of the vocal frequency.

F. Corpus

The method has been applied to several corpora of synthetic vowels [a] and to a corpus of French vowels [a] sustained by a speaker in the framework of a voice loading task. The latter involves studying the effects of vocal fatigue. Synthetic vowels [a] have been generated with different vocal jitter, vocal tremor and additive noise characteristics, with a view to validating the present approach.

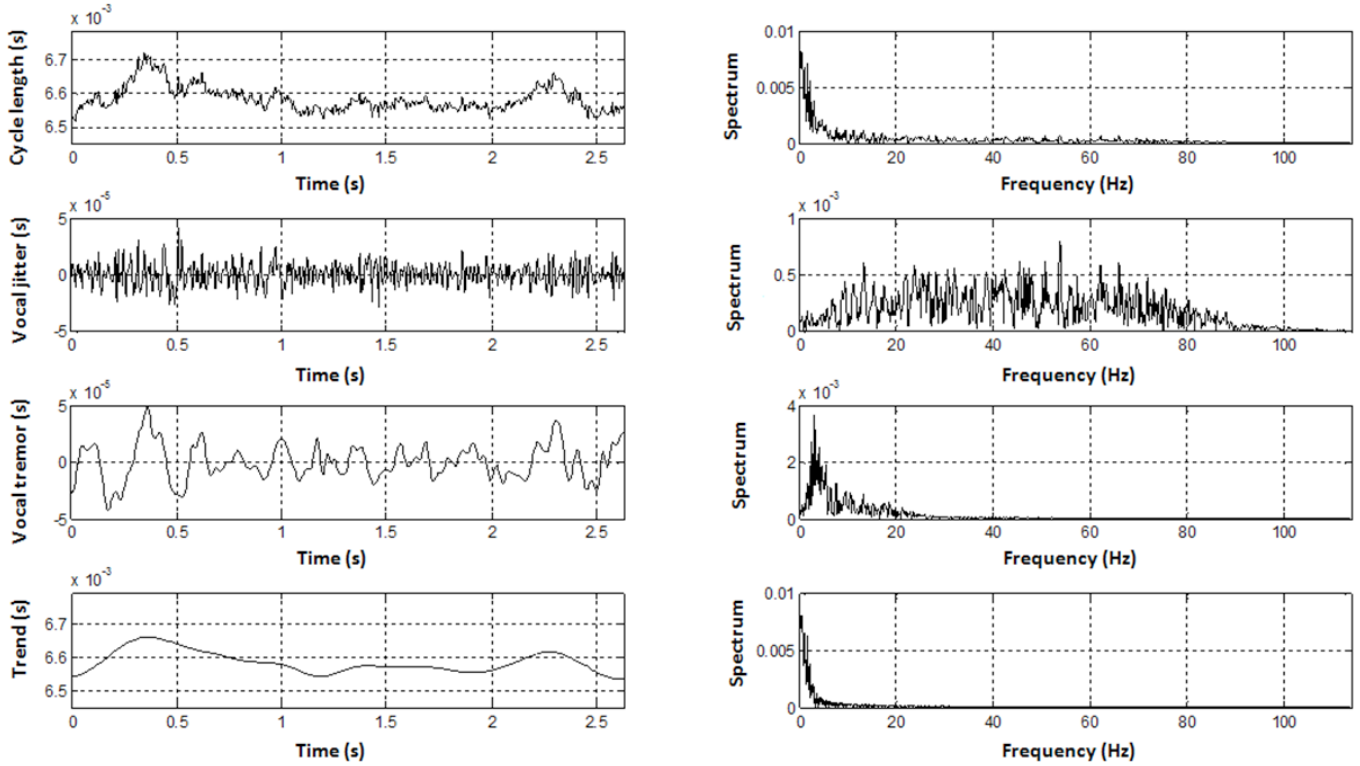


Fig. 2 : Cycle length time series and its decomposition into vocal jitter, vocal tremor and trend (temporal and frequency domain)

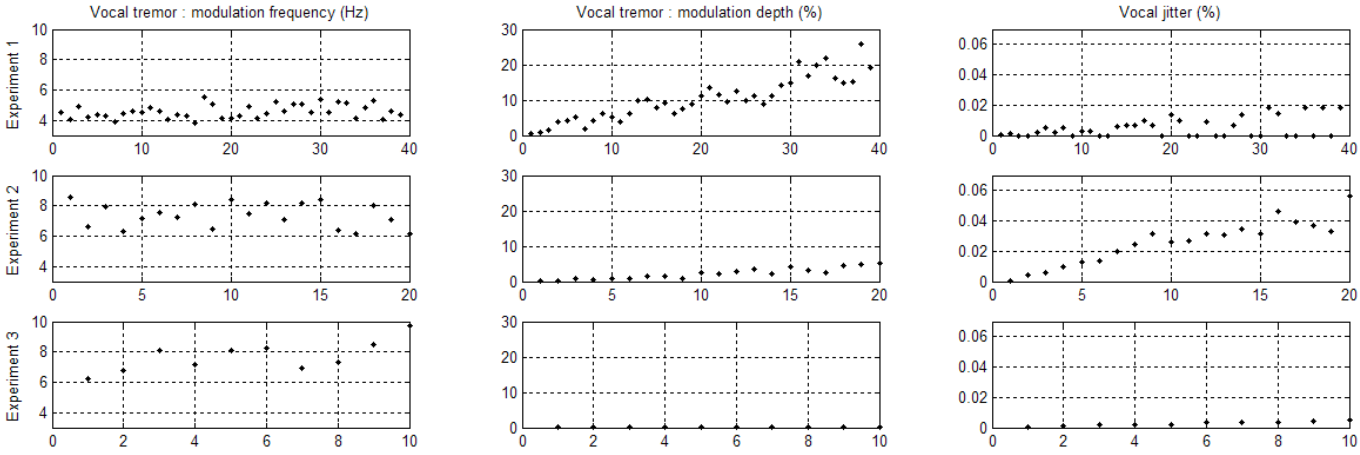


Fig. 3 : Evolution of vocal cues with different vocal tremor, vocal jitter and additive noise characteristics

III. Results and Discussion

A. Validation

The reliability of the tracking of the cycle length time series via speech peak saliences and regularity constraints had already been tested by means of modal speech signals, their numerical derivatives and integrals as well as the co-recorded throat microphone signals in [4]. Here, the reliability of the vocal cycle perturbation extraction has been tested by means of synthetic vowels, generated with different vocal tremor amplitudes (Experiment 1), vocal jitter levels (Experiment 2) and additive noise levels (Experiment 3). The default vocal tremor frequency is fixed at 4Hz. Fig. 3 shows the results of these experiments. One observes, for experiments (1) and (2), that the coefficients of variation of the tremor time series increase with the modulation depth and the coefficients of variation of the jitter time series increase with vocal jitter.

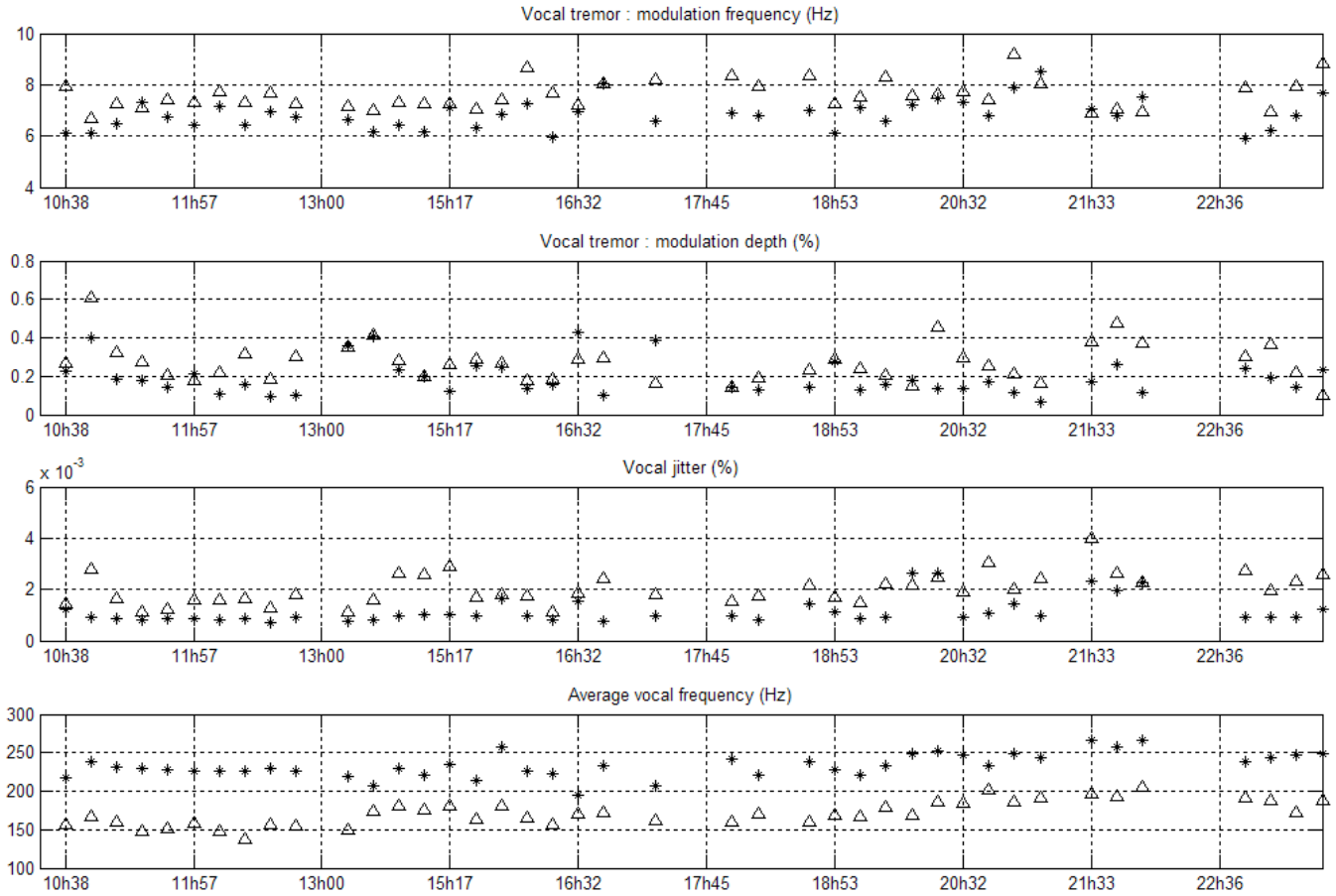


Fig. 4 : Effects over time of vocal fatigue

B. Vocal Fatigue

The results for voice loading are reported in Fig. 4. Symbol △ reports time series obtained for vowels [a] sustained at a modal vocal frequency and symbol ∗ reports results obtained for vowels [a] sustained at the highest possible vocal frequency. One observes that the effects over time of vocal fatigue are mainly an increasing vocal frequency, a possibly decreasing tremor modulation depth and an possibly increasing cycle length jitter. One also observes that tremor frequency appears not to be affected by fatigue.

REFERENCES

[1] C.Ferrer, D.Torres, and M.E.Hernandez-Diaz, "Using dynamic time warping of T0 contours in the evaluation of cycle-to-cycle pitch detection algorithms," *Pattern Recognition Letters, Elsevier*, vol. 31, pp. 517–522, 2010.

[2] N. E.Huang, Z. Shen, S. R.Long, M. C.Wu, H. H.Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H.Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceeding of the The Royal Society*, 1998.

[3] C.Mertens, F.Grenez, and J.Schoentgen, "Speech sample salience analysis for speech cycle detection," in *Proceedings 10th Annual Conference of the International Speech Communication Association INTERSPEECH, Brighton (U.K.)*, 2009.

[4] C.Mertens, F.Grenez, L.Crevier-Buchman, and J.Schoentgen, "Reliable tracking based on speech sample salience of vocal cycle length perturbations," in *Proceedings 11th Annual Conference of the International Speech Communication Association INTERSPEECH, Makuhari (Japan)*, 2010.

CEPSTRAL ANALYSIS OF PERCEPTUALLY RATED SYNTHETIC DISORDERED SPEECH STIMULI

A. Alpan[1], F. Grenez[1], J. Schoentgen[1, 2]

[1] Laboratoires d'Images, Signaux et Dispositifs de Télécommunications,
Université Libre de Bruxelles, Brussels, Belgium
[2] National Fund for Scientific Research, Belgium

Abstract: **A number of studies have shown that the amplitude of the first rahmonic peak (R1) in the cepstrum may indicate hoarse voice quality. The cepstrum is obtained by taking the inverse Fourier transform of the log-magnitude spectrum. The goal of the article is to apply cepstral analysis to a perceptually evaluated corpus of synthetic stimuli to learn about the link between the signal properties (fixed by the synthesizer parameters) and the first rahmonic peak. The synthetic stimuli have been generated by a synthesizer of disordered voices that has been shown to generate natural-sounding speech fragments comprising different vocal perturbations. A second objective is to examine the link between first rahmonic peak and perceived breathiness and roughness, link which has not been studied previously. The speech stimuli have been perceptually assessed by nine listeners according to grade, breathiness and roughness. A number of cepstral analysis alternatives have been implemented, including period-synchronous temporal frames and harmonic-synchronous band-limited analyses.**

Keywords : **cepstral analysis, synthetic disordered speech, first rahmonic amplitude**

I. INTRODUCTION

Acoustic analysis has a central place within the context of the assessment of laryngeal function because the speech signal may be recorded non-invasively and it is the basis on which the perceptual assessment of voice is founded. Many voice disorders cause voiced speech to deviate from periodicity. Dysperiodicities may be caused by additive noise owing to turbulent airflow and modulation noise owing to perturbations of the frequency and amplitude of the glottal excitation signal. Dysperiodicities may also be due to intrinsically irregular dynamics of the vocal folds and involuntary transients between dynamic regimes [1].

Several acoustic features that have been used to assess vocal fold function report the deviation of the voiced speech waveform from perfect periodicity. Vocal jitter and shimmer, for instance, are frequently used to summarize perturbations of the voiced speech cycle lengths and amplitudes, respectively. A signal that has shown promise as a global descriptor of voice quality is the cepstrum. Global descriptors designate features that report different voice qualities as patterns rather than focus on narrowly-defined properties of the speech signal.

The cepstrum is defined as the inverse magnitude spectrum of the log-magnitude spectrum [2]. Because the logarithmic power spectrum of voiced speech consists of equally spaced harmonics, a peak occurs in the inverse Fourier transform of this signal (the cepstrum) at a point corresponding to the glottal period [3]. Previous studies have shown that the amplitude of the first rahmonic peak in the cepstrum (R1) is a global descriptor of glottal turbulence noise and modulation noise [4].

Although, it has been frequently observed that increased levels of noise and perturbations in the voice signal decrease R1, a formal description of cepstral peak R1 has been lacking. Murphy has provided a theoretical description of cepstral analysis of voiced speech with aspiration noise, which suggests that R1 is directly proportional to a geometric-mean harmonics-to-noise ratio [5]. He shows that R1 and the geometric-mean harmonics-to-noise ratio (measured spectrally) underestimate the actual geometric-mean harmonics-to-noise ratio when averaged noise levels exceed harmonic levels. Limiting the number of harmonics in the analysis window overcomes this problem and in the case of period-synchronous analysis also alleviates the dependence of R1 on (temporal) window length and F0.

For the present study, a corpus of synthetic sound stimuli has been obtained by means of a synthesizer of disordered voices [6]. It can mimic a wide range of speech source perturbations such as additive noise at the glottis, vocal frequency jitter, vocal shimmer, vocal frequency tremor, amplitude tremor, diplophonia, biphonation and random glottal cycles.

The synthetic stimuli are vowels [a], [i], [u] and transients [ai] and [ia]. Each has been perceptually assessed by nine professional listeners according to grade, roughness and breathiness.

The purpose of the article is to apply cepstral analysis to a perceptually evaluated corpus of synthetic stimuli to learn about the link between the signal properties (fixed by the synthesizer parameters) and the first rahmonic peak. A second objective is to examine the link between first rahmonic peak and perceived breathiness and

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

roughness, link which has not been studied previously. A number of spectral analysis alternatives have been implemented, including period-synchronous temporal frames and harmonic-synchronous band-limited analyses.

II. Corpus

The synthesizer involves models of the glottal area and airflow through the glottis. The time-evolving glottal area is modelled by means of a nonlinear memoryless signal model that transforms a trigonometric driving function into the desired glottal area waveform. One attractive property of the model is that the frequency and harmonic richness of the glottal area are controlled by the instantaneous frequency and amplitude of the harmonic driving function.

The glottal airflow rate is generated by means of an algebraic aerodynamic model involving the glottal area and including interactions between the glottis and the infra- and supra-glottal ducts. The propagation of the acoustic wave through the trachea and vocal tract is simulated by means of concatenated tubes. Wall vibration, viscous and thermal losses as well as acoustic reflection and radiation at the lips and glottis are taken into account. Modulation noise such as jitter or tremor and abnormal voice qualities such as diplophonia, biphonation and irregular vocal cycles are mimicked by means of stochastic or deterministic models of the time-evolving instantaneous frequency or amplitude of the driving harmonics of the glottal area model.

The ability of the synthesizer to mimic natural disordered voices has been demonstrated in the framework of several perceptual experiments [7].

The corpus comprises synthetic sounds [a], [i], [u], [ia] and [ai] which are one second long. The vowel → vowel transitions have been simulated by evolving linearly the tract area function from one vowel target to the next over an interval of 0.2 s in the middle of the one second interval. Each set is composed of 48 stimuli that combine three values of vocal frequency, four levels of frequency jitter and four levels of additive noise. The vocal frequency values are 100, 120 and 140 Hz. The jitter and additive noise have been fixed based on the independent advice of one phoniatrician and one speech therapist so that the stimuli are perceived as covering the full ranges of grade (G0 - G3), roughness (R0 - R3) and breathiness (B0 - B3) on the GRB(AS) scales. The area function of the vocal tract has been identical for all stimuli of the same vowel category [6].

Eight speech therapists and one phoniatrician have perceptually evaluated each set of synthetic sounds according to perceived “grade” (G), “roughness” (R) and “breathiness” (B) with four degrees per scale: 0 (normal), 1 (feeble), 2 (moderate) and 3 (severe). The scores of G, R and B have been averaged over the nine judges.

III. Methods

A. First rahmonic amplitude (R1)

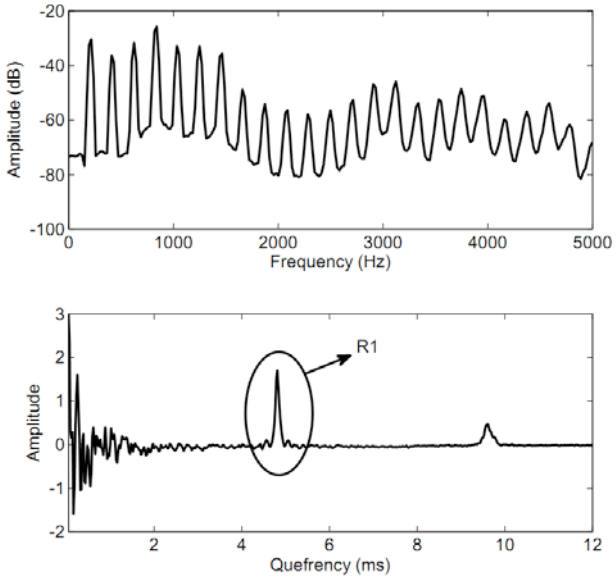


Figure 1: Amplitude spectrum and cepstrum of sustained vowel [a] showing the first rahmonic amplitude.

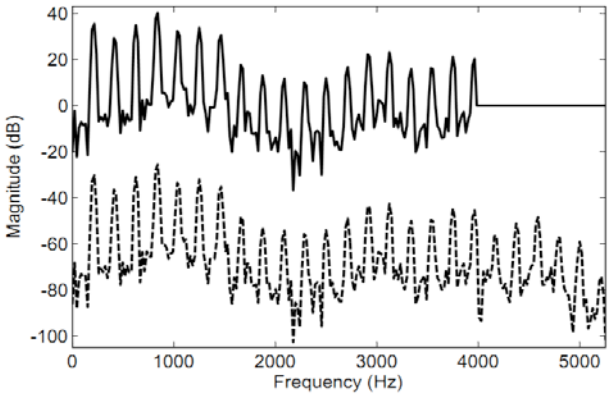


Figure 2: Dashed line: Log-magnitude spectrum. Solid line: Band-limited and offset-removed log-magnitude spectrum for an analysis frame of vowel [a].

- Full-band spectrum

The computation of the amplitude of first rahmonic (R1) involves the following. The amplitude spectra of the hopped Hamming-windowed frames are averaged and the log-magnitude of the average is taken (Fig. 1). The cepstrum is obtained via the inverse Fourier transform of the log-amplitude average spectrum. First rahmonic R1 is located in the vicinity of the quefrency corresponding to the glottal cycle length. The analysis has been period-synchronous, i.e. the lengths of the frame have been multiples of the vocal cycle length.

- Band-limited spectrum

The computation of R1 implicates the same steps as previously. However, prior to the computation of the cepstrum, the log-average spectrum is limited to a fixed number of harmonics and the offset is removed (Fig.2).

B. Correlation analysis

The correlation coefficients of the amplitude of first rahmonic (R1) obtained via the different options (full-band, band-limited) with the average perceptual scores for grade, roughness and breathiness have been computed.

C. Multi-cue regression analysis

The amplitude of first rahmonic (R1) has been regressed on the parameters of the synthesizer fixing additive noise, jitter and fundamental frequency. The phonetic category of the stimuli has been taken into account by a dummy variable: 1 for [a], 2 for [ai], 3 for [ia], 4 for [i] and 5 for [u].

IV. RESULTS

A. Correlation analysis

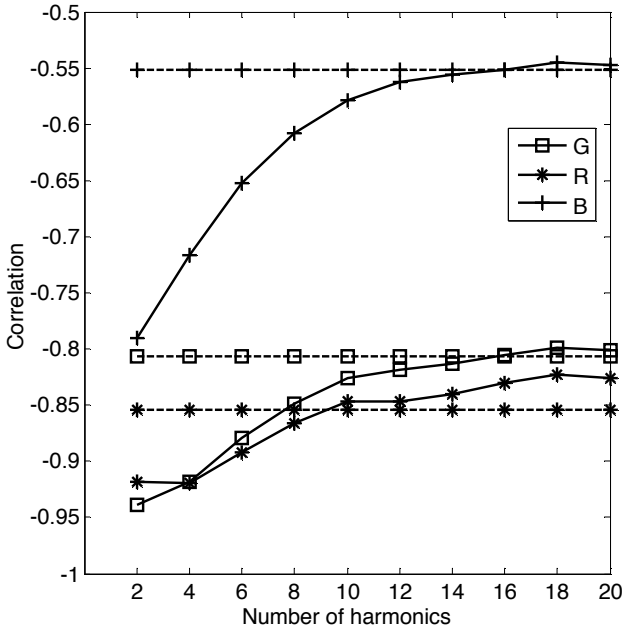


Figure 3: Correlation of period-synchronous (6 cycles) band-limited R1 with perceptual scores GRB of vowel [a]. The dashed lines correspond to the correlations of full-band R1 with GRB.

Fig. 3 displays the correlations of period-synchronous (6 periods) full-band and band-limited R1 with average perceptual scores for vowel [a]. One observes that R1 is highly correlated with the average scores of roughness and grade and moderately correlated with breathiness. Limiting the spectrum to a feeble numbers of harmonics prior to computing the cepstrum enables improving the correlation with average perceptual scores. In particular, one observes that if the spectrum is limited to only two harmonics a correlation of 0.80 with the average breathiness scores is obtained. This correlation rapidly decreases to 0.55 when increasing the number of harmonics in the spectrum.

The correlations obtained for vowel [i] are slightly smaller but similar to other vowels and vowel-vowel pairs.

B. Multi-cue regression analysis

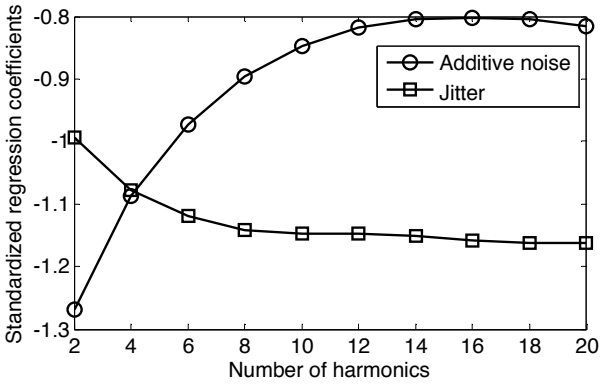


Figure 4 : Standardized regression coefficients of the additive noise and jitter control parameters predicting R1.

The period-synchronous (6 periods) band-limited R1 has been predicted via the additive noise, jitter and vocal frequency control parameters of the synthesizer and a dummy variable that takes phonetic category into account. Fig. 4 displays the standardized regression coefficients of the additive noise and jitter parameters. Vocal frequency and phonetic category do not contribute to the prediction of period-synchronous R1. Jitter contributes most. However, when the number of harmonics in the band-limited spectrum is smaller than 6, the contribution of pulsatile additive noise exceeds the contribution of jitter. Also, when the number of harmonics in the spectrum increases the correlation of R1 with additive noise decreases (down to 0.35) while the correlation with jitter increases (to 0.85).

V. DISCUSSION AND CONCLUSION

Correlations of full-band R1 with perceived roughness ($\rho \approx 0.85$) and grade ($\rho \approx 0.80$) are good. However, only a moderate correlation is observed with perceived breathiness ($\rho \approx 0.55$).

Limiting the spectrum to a feeble number of harmonics improves the correlation. The largest

improvement is observed for breathiness. Indeed, when the spectrum is limited to 2 harmonics, the correlation increases to 0.80.

One has also observed that jitter contributes most to predicting the period-synchronous band-limited R1 (Fig.4) when the number of harmonics in the spectrum is larger than 6. A possible explanation is that cue R1 mainly reports modulation noise that broadens and decreases harmonic amplitudes and adds spectral sidebands. In the case of homogeneous modulation noise, spectral effects are proportional to the order of the harmonic. When the number of harmonics that is taken into account is feeble at low-frequencies, modulation noise effects are less prominent in R1 and the influence of additive noise, which is harmonic independent, increases. Fig. 5 displays the average spectrum for two vowels [a] with the same additive noise level and low and high vocal jitter levels.

One observes that the correlation of roughness and grade with cue R1 increases (Fig.3) whereas the correlation of R1 with jitter decreases (Fig.4) when the bandwidth of the log-amplitude spectrum decreases. A possible explanation is that experiments reported elsewhere suggest that the perception of roughness is effected both by modulation noise and additive pulsatile noise.

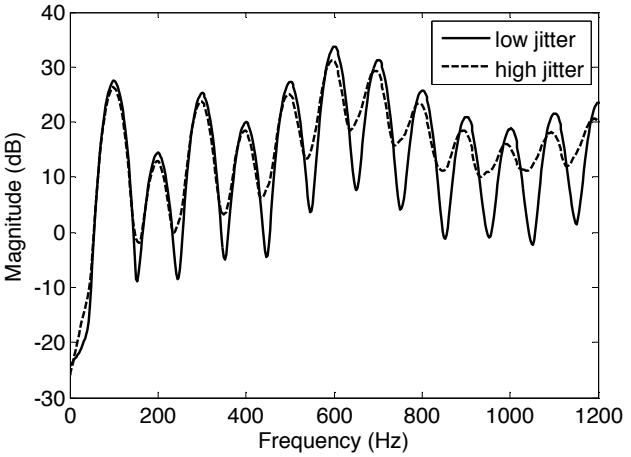


Figure 5 : Average spectrum for two vowels [a] with low (solid line) and high (dashed line) vocal jitter level.

Some of these results may seem to contradict simulations reported in [5], which have shown that period-synchronous band-limited R1 evolves linearly with aspiration noise and non-linearly with jitter.

The main reason for the differences between both studies is that in [5] experiments have been carried out on stimuli perturbed by one kind of noise only. Therefore, the effects of one type of perturbation are not masked by another type. In particular, the spectral effects of additive noise are not hidden by those of vocal jitter.

Additional experiments have therefore been carried out on synthetic disordered stimuli [8] resting on the synthesis model reported above, but involving one type of noise only.

One then observes that period-synchronous (6 periods) harmonic-limited R1 obtained from stimuli perturbed by the additive noise only is very well correlated with the noise level ($\rho \approx 0.90$). This correlation is larger than the correlation obtained for stimuli containing vocal jitter only ($\rho \approx 0.80$).

Also, the synthetic stimuli used here are more natural than in [5]. Indeed, in [5] the purpose of the stimuli was not to mimic natural disordered voice and the noise characteristics also differ between both studies. Aspiration noise in [5] was synthesized by means of a zero-mean white Gaussian noise added to the glottal source, whereas in this study, additive noise is mimicked by means of Brownian noise, the amplitude of which is modulated via an affine function of the glottal airflow rate [7]. Also, in [5] vocal jitter is synthesized through time scaling of glottal waveforms, whereas here it is caused by small random perturbations of the instantaneous frequency of the driving function of the synthesizer.

VI. Acknowledgements

This research has been supported by COST ACTION 2103 “Advanced Voice Function Assessment” in the framework of a short-term scientific mission at the University of Limerick, and by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme.

References

[1] J. Schoentgen, Spectral models of additive and modulation noise in speech and phonatory excitation signals. J. Acoust. Soc. Am. 113, 2003, pp. 553-562.

[2] A.V. Oppenheim, R.W. Schafer, “Digital Signal Processing,” Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

[3] L. Rabiner, R. Schafer, “Digital Processing of Speech Signals,” Prentice–Hall, Englewood Cliffs, New Jersey, 1978.

[4] S. N. Awan, N. Roy, “Outcomes measurement in voice disorders: Application of an acoustic index of dysphonia severity,” Journal of Speech, Language, and Hearing Research 52, 2009, pp. 482-499.

[5] P.J. Murphy, “On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals,” J. Acoust. Soc. Am., 120 (5), 2006, pp 2896–2907.

[6] S. Fraj, “Synthèse des voix pathologiques”, PhD thesis, Université Libre de Bruxelles, 2010.

[7] S. Fraj, F. Grenez and J. Schoentgen, “Perceived naturalness of a synthesizer of disordered voices”, in Proc. Interspeech, Brighton (U.K.), 2009, pp. 1178-1181.

[8] J. Hanquinet, F. Grenez and J. Schoentgen, “Synthesis of disordered speech,” In Proc INTERSPEECH 2005, Lisbon, Portugal, September 2005, pp.1077-1080.

SYNTHESIS OF BREATHY AND ROUGH VOICES WITH A VIEW TO VALIDATING PERCEPTUAL AND AUTOMATIC GLOTTAL CYCLE PATTERN RECOGNITION

S. Fraj[1], F. Grenez[1], J. Schoentgen [1,2]

[1] L.I.S.T, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium
[2] National Fund for Scientific Research, Belgium

Abstract: **The framework of the presentation is the assessment of the ability of human raters or speech-processing software to detect glottal cycles in speech sounds and measure their lengths in synthetic breathy and rough voices. The synthesis of hoarse voices designates the generation of speech sounds the timbre of which simulates the voice quality of dysphonic speakers. The added value of synthetically generated test stimuli is that the user may fix and know their properties exactly. The corpus comprises synthetic vowels [a] combining seven levels of frequency jitter and three levels of additive noise. The presentation is focused on the simulation of rough and breathy voices via frequency modulation of the glottal excitation model and addition of pulsatile noise at the glottis. Furthermore, the genuine glottal cycle lengths and glottal source to noise ratios are obtained to which lengths and ratios inferred via signal processing may be compared. The glottal cycle lengths are acquired by tracking the phase of the harmonic driving functions of the speech sound synthesizer. Actual glottal signal-to-noise ratios are measured by summing separately over the sound stimuli the squared clean volume velocity and pulsatile noise samples.**

Keywords: **speech synthesis, breathiness, roughness, frequency jitter, amplitude shimmy, and additive glottal noise.**

I. INTRODUCTION

The synthesis of breathy or rough voices designates the generation of speech sounds the timbre of which simulates the voice quality of dysphonic speakers. The framework of the present article is the assessment of the ability of human raters or speech-processing software to detect and measure the length of glottal cycles in severely hoarse voices. The added value of synthetically generated test stimuli is that the user may fix and know their properties exactly. Artificial speech sounds have therefore been used in the past to develop and test clinical analysis software the purpose of which is to obtain cues that describe a speaker's voice quantitatively. A distinction between synthetic and artificial speech sounds may be apposite, however. Indeed, synthetic speech, whatever its purpose, is meant to be listened to by humans to whom it must sound intelligible and natural. The property of naturalness may be relevant because it causes synthetic speech to be perceived as human or nearly human so that a listener's or a speech processing software's response to the paralinguistic or extra-linguistic (e.g. clinical) facets of the synthetic speech sounds are similar to their responses to genuine human stimuli

II. METHODS

A. Synthesis of disordered speech sounds

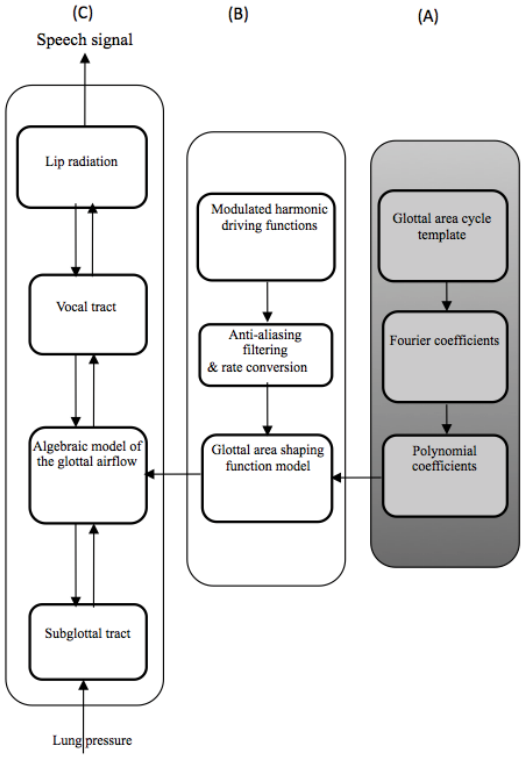


Fig.1: Scheme of the synthesizer

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

The synthesis involves three modules (A) – (C) that feed each other from right to left in *Fig*. 1. The right-most module (A) is only active for initialization. Modules (B) and (C) are active during synthesis. The purpose of module (A) is the computation of the coefficients of distortion polynomials by means of which the glottal area is simulated [2]. The polynomial coefficients are obtained via a constant linear transform from the Fourier series coefficients of the template cycle, which has been a Klatt model of the glottal area [5].

The generation of signal cycles via polynomials and their control by means of the parameters of a driving sinusoid is known as nonlinear wave shaping. Wave shaping enables setting sample by sample the frequency of the glottal area and continuously evolving its shape between a quasi-sinusoid and the default template shape that is fixed via the coefficients of the wave-shaping polynomials.

During synthesis, Module (B) simulates the glottal area based on a nonlinear transform of two harmonic functions via two distortion polynomials into the desired area waveform. The instantaneous frequencies of the harmonic driving functions are identical and they fix the frequency of the glottal area waveform. The identical amplitudes of the driving functions fix the open quotient and spectral brilliance (bandwidth) of the glottal area the overall amplitude of which is determined by a linear gain. When the driving amplitudes are equal to one, the nonlinear distortion model outputs the default shape and size of the Klatt area template. When the driving amplitude is small, the output of the wave-shaper is quasi-sinusoidal and a constant when the driving amplitude is zero. At the top of module (B) the sampling rate is equal to 176 *kHz* and 88 *kHz* at the bottom.

In module (C), trachea and vocal tract are imitated by concatenations of short elementary cylinders. The number of ducts of the trachea is equal to 36. Their constant cross-section is equal to 1.2 cm^2. The number of elementary cylinders mimicking the vocal tract is comprised between 40 and 45, depending on the vowel. The cylinder cross-sections have been fixed on the base of published data [7] [8] [9]. The acoustic wave generated at the glottis by the glottal volume velocity propagates through the trachea, where part of the wave is absorbed by the lungs, and vocal tract, where part of the wave is radiated at the lips giving rise to audible sound [10]. In addition, the model involves simulations of the acoustic interactions at the glottis of the trachea and vocal tract as well as of the losses owing to friction at and vibration of the tract walls, and also to passing heat through the walls [11][12]. The glottal volume velocity is simulated by means of an algebraic model of the glottal aerodynamics that is driven by the glottal area waveform and that takes into account sub and supra-glottal pressures to include effects of source-tract interactions [1]. *Fig*. 2 shows two cycles of the glottal area waveform and the corresponding volume velocity waveform.

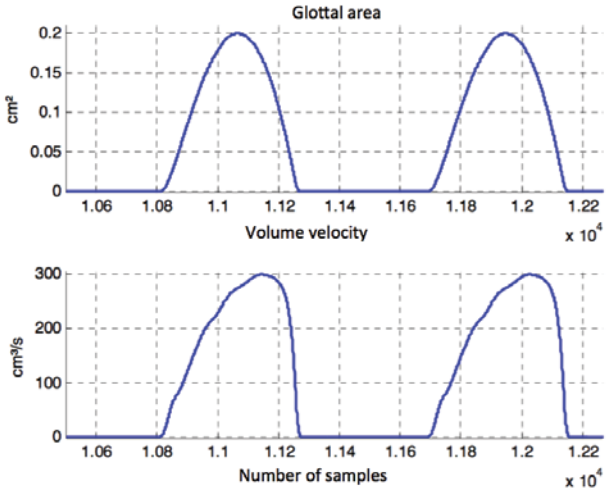


Fig.2: Glottal area and volume velocity as a function of time in number of samples. Top: glottal area, bottom: volume velocity.

B. Simulation of voice disorders

The simulations of voice disorder involve simple formal models of frequency modulation noise and additive noise owing to turbulence. Modulation noise models determine the driving function parameters at the top of module (B) and turbulence noise models co-determine the volume velocity generated at the glottis. The synthetic stimuli that have been prepared for the experiment that are reported in other presentations of the same session differ by their level of frequency jitter and level of additive pulsatile noise. The default (unperturbed) vocal frequency has been fixed at 100 Hz. Unusual glottal dynamics such as diplophonia or biphonation are not involved.

Frequency jitter is simulated by means of small positive and negative noise samples that disturb the instantaneous phase Θ of the harmonic driving functions that generate the glottal area as shown in *Fig*. 1. The size of the noise samples ξ is fixed by means of parameter b in formula (1) and their sign (plus or minus) is assigned stochastically with equal probability p. The small sample-by-sample phase disturbances add up over one glottal cycle to the observed cycle length perturbations known as jitter [3]. They also cause shape perturbations of the glottal area, which are minor when the instantaneous perturbations are modal. Symbol f_0 is the unperturbed vocal frequency and Δ is the sampling step. *Fig*. 3 shows an example of an extremely hoarse voice (b = 4.5, n_1 = 0.55) the purpose of which has been to test human as well as vocal cycle pattern recognition by machine.

$$\theta(n) = \theta(n-1) + 2\pi f_0 \Delta + 2\pi b \xi(n)$$
$$\xi(n) = \begin{cases} +1, p = 0.5 \\ -1, p = 0.5 \end{cases} \quad (1)$$

Amplitude perturbations (vocal shimmy) of the speech cycles are generated via modulation distortion by the vocal tract transfer function that turns frequency perturbations of the glottal volume velocity into amplitude perturbations of the speech signal [4]. Intuitively speaking, cycle amplitude perturbations are the combined effect of perturbations of the volume velocity harmonic amplitudes owing to perturbations of the harmonic positions on the frequency axis. The amplitudes of the harmonics indeed change when their positions change because the vocal tract transfer function is not flat.

Breathiness has been simulated by pulsatile additive noise, which simulates additive noise owing to turbulent airflow in the vicinity of the glottis. Additive pulsatile noise here designates noise the size of which evolves proportionally to the glottal volume velocity. Aspiration noise designates audible noise owing to constant airflow through a static glottis or permanent glottal chink. The reason pulsatile noise has been added to the clean glottal volume velocity instead of static aspiration noise is that strong stationary glottal noise and voice segregate into two distinct auditory streams [6].

Glottal noise (2) is simulated by means of low-pass filtered white Gaussian noise the standard deviation of which is fixed and the samples of which are multiplied by the clean glottal volume velocity u_g and delayed by 1 *ms* before they are added to u_g.

$$n_1 u_g(n) + n_2 \quad (2)$$

The user selects the value of coefficient n_1, which fixes the amount of pulsatile noise, and a constant offset n_2 that mimics aspiration noise, which is small, compared to the pulsatile component.

C. Tracking of glottal cycle length perturbations and glottal source to noise ratios

Formulas (1) and (2) show that the simulation of modulation and additive noise involves stochastic components that impede predicting the amount of cycle-to-cycle perturbations as well as the signal-to-noise ratio exactly. Also, synthetic cycle length perturbations and glottal turbulence noise are the output of models that are dependent on several parameters. These quantities, therefore, cannot be controlled manually. As a consequence, the reference quantities to which the user-observed cycle lengths and glottal signal to noise ratios may be compared must be traced synthesizer-internally because they are neither directly controllable nor observable from the outside.

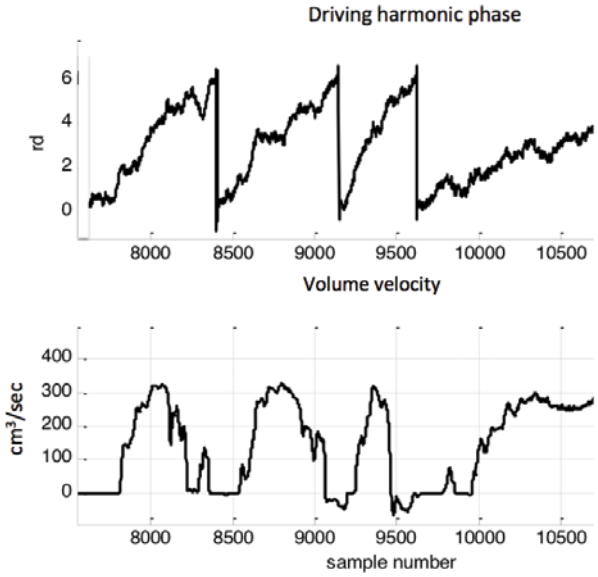


Fig.3: Glottal cycles of an extremely hoarse voice. Top: Phase of harmonic driving functions; bottom: volume velocity

The glottal cycle lengths have been obtained by tracking the phase of the harmonic driving functions. The onset of each glottal cycle is assigned to the time sample when the phase is reset from 2π to zero. The distance between two successive onsets obtains the genuine glottal cycle length in number of samples. Figure 3 shows a few cycles of the perturbed harmonic driving function phase (top) and perturbed and noisy volume velocity (bottom).

The squares of the clean volume velocity and noise samples have been summed separately over the signal length. The log-ratio of both sums multiplied by ten is the signal-to-noise ratio in dB at the glottis. Even for signals rated as very breathy, this ratio is high (> 17 dB). The explanations are that speech sound radiation at the lips favors high over low frequencies, that is, the perceived noise is broadband compared to the glottal noise and the bandwidth of the noise is anyway larger than the bandwidth of the clean glottal volume velocity because the spectral slope of the latter is steeper. Also, the noise is pulsatile, that is, strong over part of the cycle and negligible over the rest. This means that the glottal noise energy averaged over the signal length is a flawed predictor of perceived breathiness.

III. Results

Table 1 summarizes a corpus of 21 two-second long sustained vowels [a] with gradual on and offsets. Each synthetic voice has a default frequency of 100 Hz and typically comprises 200 cycles. The corpus is a combination of seven (extreme) levels of cycle length jitter, corresponding to values of parameter *b* in formula (1) increasing from 0.315 to 4.5 and of three increasing levels of additive low-pass filtered white Gaussian noise, corresponding to values of parameter n_1 equal to 0.15, 0.35 and 0.55. The values of parameter n_1 are fixed to cause B1, B2, B3 breathiness scores typical of disordered voices. The Table reports the relative cycle length jitter at the glottis in % and the volume velocity energy to noise energy log-ratio in dB corresponding to the seven jitter levels and three noise levels mentioned above. These data are obtained by tracking cycle lengths and noise synthesizer-internally as described in section *C*.

IV. Discussion and conclusion

The performance of speech processing software as well as human raters when tested by means of this corpus is reported in separate presentations.

References

[1] Titze, I. (2006). "The myoelastic aerodynamic theory of phonation", (National Center for Voice and Speech, Denver CO, Iowa City IA).
[2] Schoentgen, J. (2003a). "Shaping function models of the phonatory excitation signal", J. Acoust. Soc. Am., 114, 2906–2912.
[3] Schoentgen, J. (2001). "Stochastic models of jitter", J. Acoust. Soc. Am.,109, 1631–1650.
[4] Schoentgen, J. (2003b). "Spectral models of additive and modulation noise in speech and phonatory excitation signals", J. Acoustic. Soc. Am., 113, 553–562
[5] Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer", J. Acoustic. Soc. Am. 67, 971–995.
[6] Yiu, E. M.-L., Murdoch, B., Hird, K., and Lau, P. (2002). "Perception of synthesized voice quality in connected speech by cantonese speakers", J. Acoust. Soc. Am. 112, 1091–1101.
[7] George, M. (1997). "Analyse du signal de parole par modélisation de la cinématique de la fonction d'aire du conduit vocal", Ph.D. thesis, Faculty of Sciences, Université Libre de Bruxelles.
[8] Story, B. H., Titze, I. R., and Hoffman, E. A. (1996). "Vocal tract area functions from magnetic resonance imaging", J. Acoust. Soc. Am. 100, 537–554.
[9] Mryayti, M. (1976). "Contributions aux ´etudes sur la parole, Thèse d'Etat", Institut National Polytechnique de Grenoble, France.
[10] Flanagan, J. L. and Rabiner, L. R. (1972). "Speech synthesis", Bell Syst. Technol. J. 1233–1268.
[11] Hanquinet, J., Grenez, F., and Schoentgen, J. (2005). "Synthesis of disordered speech", in Proc. INTERSPEECH 2005, 1077–1080 (Lissabon, Portugal).
[12] Abel, J., Smyth, T., and III, J. O. S. (2003). "A simple, accurate wall loss filter for acoustic tubes", in Proc. 6th Int. Conference on DigitalAudio Effects (DAFx-03) (London, UK).

Tab.1: Relative cycle length jitter at the glottis in % and the volume velocity energy to noise energy log-ratio in dB for different values of parameters b and n_1

b	n_1	Glottal cycle length jitter (%)	Volume velocity to noise ratio (dB)
0.315	0.15	2.6	28.8
0.315	0.35	2.5	23.5
0.315	0.55	2.6	17.5
0.63	0.15	4.5	30.0
0.63	0.35	5.3	22.7
0.63	0.55	5.6	18.5
1.26	0.15	9.5	29.4
1.26	0.35	8.8	21.7
1.26	0.55	10.0	17.4
1.89	0.15	14.7	28.8
1.89	0.35	14.7	22.9
1.89	0.55	15.7	19.2
2.52	0.15	21.9	29.4
2.52	0.35	20.4	22.2
2.52	0.55	20.9	18.6
3.45	0.15	24.4	29.6
3.45	0.35	24.1	22.7
3.45	0.55	27.2	17.7
4.5	0.15	31.4	29.2
4.5	0.35	31.6	22.0
4.5	0.55	35.8	18.5

TREMOR IN SPEAKERS WITH SPASMODIC DYSPHONIA

Maria Koutsogiannaki, Yannis Pantazis, Yannis Stylianou[1] and Philippe Dejonckere[2]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece

[2]Utrecht University, Utrecht, The Netherlands

email: {mkoutsog, pantazis, yannis}@csd.uoc.gr, Philippe.Dejonckere@fmp-fbz.fgov.be

***Abstract* - The objective of this work is the estimation of vocal tremor in patients with spasmodic dysphonia before and after treatment, and the comparison of their tremor characteristics with those estimated from healthy speakers. As an outcome, a new tremor attribute is introduced, the deviation of the modulation level and a novel method is proposed for classifying speakers according to the prevalence of tremor in their voice. Results are consistent with subjective evaluations on patients who suffer from spasmodic dysphonia and confirm that the proposed method can be used for accurate estimation and objective ranking of the severity of tremor.**

***Index Terms*—Voice quality, vocal tremor, dysphonia, deviation of modulation level**

I. INTRODUCTION

Vocal tremor, a rhythmic change in pitch and loudness, appears both in healthy speakers and in speakers with voice disorders. In normal speaking voice, no tremor is audible, but it can be elicited by emotions, either spontaneous or volitional (actors). Central (mostly degenerative) neurological diseases, particularly those involving cerebellum and basal ganglia, frequently elicit voice tremor. In spasmodic dysphonia (or laryngeal dystonia), task-related tremor ("spasms") may considerably hamper fluency and intelligibility [1]. This work focuses on estimating tremor in speakers with spasmodic dysphonia before and after treatment, and compares their tremor characteristics (level and frequency) with those estimated from healthy speakers.

Acoustic analysis of tremor is usually based on the accurate estimation of fundamental frequency and then the characterization of the fundamental frequency's variations [2], [3]. Modulation frequency and modulation level are prominent attributes that are extracted from the instantaneous fundamental frequency [2], [3]. Previous studies in tremor analysis assume modulation frequency and modulation level being as time-invariant characteristics of tremor, by considering short-time analysis windows of speech. Then, stationary frequency estimation approaches are used for the estimation of these tremor attributes, like the classical Fourier transform. However, tremor characteristics and in general modulations in speech are time-varying. Actually, analysis of large segments of speech showed interesting time-varying characteristics on vocal tremor [4], [5].

The detection of tremor attributes in a speech signal involves the accurate extraction of the signal that modulates the time-varying fundamental frequency. We employ a recently proposed method to extract time-varying tremor attributes; the level and the frequency of the modulating signal [6]. This method is applied to sustained vowels and decomposes the speech signal into its time-varying quasi-harmonics. Quasi-harmonics are components with frequencies which are near to be harmonics of a fundamental frequency. It has been shown that speech is better modeled as a sum of quasi-harmonics rather than a sum of harmonics [7]. Next, we will refer to the components rather to harmonics. After the decomposition of speech into components, one component is chosen for further analysis; the desired signal that modulates the component is extracted and its time-varying amplitude and frequency are estimated.

This method is applied in speech vowels uttered by normophonic speakers and speakers who suffer from spasmodic dysphonia before and after imposed on medical treatment [8]. Our analysis shows that the mean modulation level in dysphonic speakers is distinguishably greater than that in normophonic speakers. However, the modulation level is not the only criterion for classifying speakers as normophonic or dysphonic. This study introduces a novel attribute of tremor which derives from the time-varying characteristic of the modulation level, namely the deviation of the modulation level. The mean modulation level and its deviation are combined in a quality indicator trying to classify speakers according to the amount of tremor in their voice. It is shown that this objective classification of speakers matches subjective evaluations by experts in the case of spasmodic dysphonia patients.

The organization of the paper is as follows. Section II describes briefly the tremor estimation method. Section III presents the analysis on normophonic and dysphonic speakers, introduces the proposed tremor classification method and compares the results with the subjective evaluations. Finally, Section IV concludes the paper.

II. ESTIMATION OF VOCAL TREMOR

The method used for tremor features estimation assumes speech as a sum of time-varying sinusoids [7], [9]. The extraction of vocal tremor characteristics is carried out in three steps, following the procedure in [6]. The first step estimates the instantaneous amplitude and instantaneous frequency of every sinusoid component of the speech signal using a recently proposed AM-FM decomposition algorithm, the so-called Adaptive Quasi-Harmonic Model (AQHM) [7], [9].

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

AQHM is an adaptive algorithm which is able to represent accurately multi-component AM-FM signals like speech. In the second step, the very slow modulations ($< 2Hz$), derived mainly from the pulsation of the heart, are subtracted from the instantaneous component. This is achieved by filtering the instantaneous component using a Savitzky-Golay smoothing filter [10]. In the final step, the time-varying modulation frequency and the time-varying modulation amplitude of the analyzed instantaneous component are estimated by employing again the AQHM algorithm for just one component. The time-varying modulation amplitude with an appropriate scaling corresponds to the modulation level. The scaling is necessary because the modulation amplitude is relative to the mean value of the instantaneous component and involves the normalization of the amplitude by this mean value. More details of the estimation algorithm are provided in [6].

III. Results

A. Data Analysis

The suggested tremor estimation method, as described in Section II, is applied to two different databases of sustained vowels to extract the time-varying modulation level and the time-varying modulation frequency. The first database consists of sixteen healthy subjects. Sustained vowels */a/*, */e/*, */i/*, */o/* and */u/* of varying duration ($2s-8s$) have been recorded. The second database was provided by the last coauthor (Prof. P. Dejonckere). Speakers in this database suffer from spasmodic dysphonia and are subjected to treatment (botulinum toxin injections). Recordings and subjective evaluations by experts have been made before and after the treatment. For every patient, the sustained vowels of */a/* are extracted to create the signals for our analysis. In the current study, five untreated speakers could not be analyzed because they could only provide phonemes with very limited duration (less than a second).

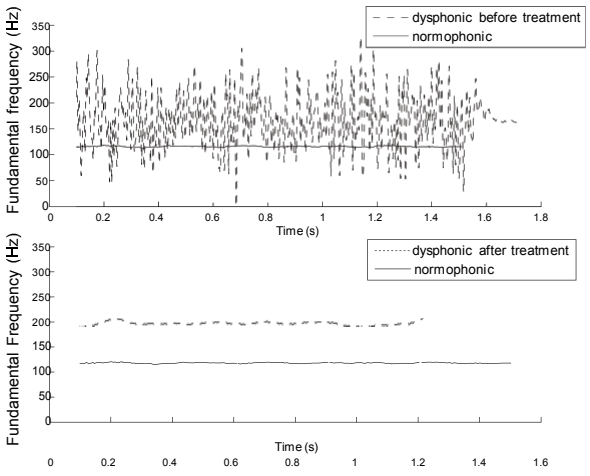


Fig. 1. The time-varying instantaneous component of a normophonic speaker and of a speaker with spasmodic dysphonia before and after treatment.

The upper panel of Fig.1 shows a typical example for the time-varying frequency characteristics of the first component (nearly the fundamental frequency) for a dysphonic and a normophonic male speaker. It is worth noticing the high fluctuations of the component for the case of the dysphonic speaker in contradiction to that of the healthy speaker who keeps his voice almost steady in time. After treatment the dysphonic speaker achieves to stabilize his voice (lower panel of Fig.1). The tremor attributes of these signals, the modulation level and the modulation frequency, are depicted in Fig.2. The upper panel of Fig.2 shows the time-varying modulation levels of a normophonic and that of a dysphonic speaker before and after his treatment. The lower panel of Fig.2 depicts the corresponding modulation frequencies. As it can been seen, the normophonic speaker appears to have much lower mean modulation level than the dysphonic speaker before treatment. Moreover, the modulation level of the dysphonic speaker before treatment presents high fluctuations over time. After treatment, both speakers have similar modulation levels; the modulation level of the treated dysphonic speaker has decreased significantly, meaning that the tremor is no longer audible after treatment. In all cases, modulation frequency values are quite comparable (lower panel in Fig.2).

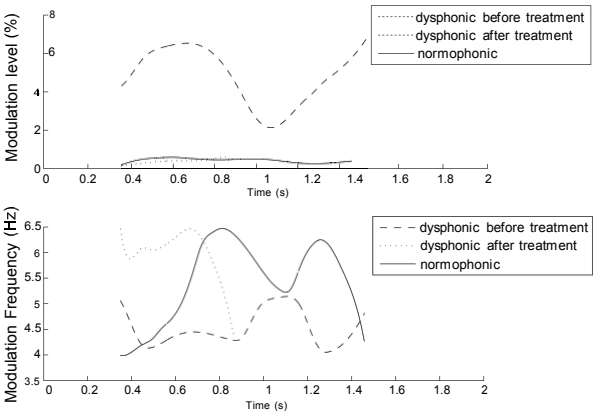


Fig. 2. Modulation level and modulation frequency of a normophonic speaker and of a speaker with spasmodic dysphonia before and after treatment.

Significant results derive from the analysis of the two databases. Fig.3 shows the mean values of the two time-varying tremor attributes for every normophonic speaker in the first database; the mean modulation level (upper panel of Fig.3) and the mean modulation frequency (lower panel of Fig.3). Frequencies vary from $2-7Hz$, while the mean modulation levels are all but one below 1% of the mean value of the instantaneous component for the corresponding normophonic speakers. In a similar way, the upper panel of Fig. 4 shows the mean modulation levels and the lower panel of Fig.4 the mean modulation frequencies for dysphonic speakers before and after their treatment. Comparing Fig.4 and Fig.3 it can be seen that the modulation frequencies are quite comparable for the normophonic and dysphonic speakers. However, this is not true for the modulation level. Indeed, five out of six untreated dysphonic speakers have modulation level above 1% and seven out of nine treated dysphonic speakers have modulation level below 1%. This is more evident in Fig.5, where the modulation level for each dysphonic speaker before

and after treatment is illustrated. For the speakers coded as Lul, Roo and Stu the modulation level has decreased after treatment, while for Bru and Vro there is a slight increase in the modulation level after the treatment. The general trend, however, is that the treated patients have modulation level values below 1% of the mean value of their component and this is comparable with that of the normophonic speakers.

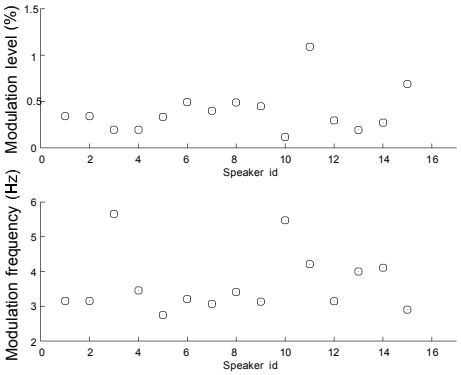


Fig. 3. Modulation levels and modulation frequencies of normophonic speakers.

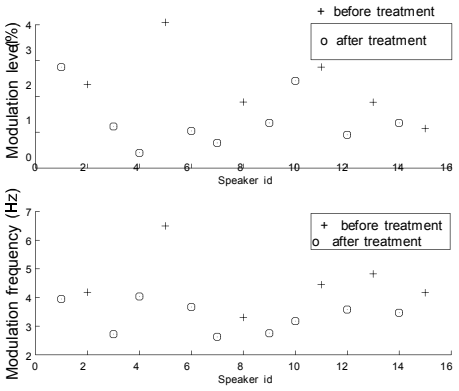


Fig. 4. Modulation levels and modulation frequencies of treated and untreated dysphonic speakers.

As illustrated by the evolution of the modulation level in the upper panel of Fig.2, the deviation of the modulation level from its mean value is quite high in the case of the dysphonic speaker before treatment. This was also observed in other dysphonic speakers from the same database. Based on this observation a new characteristic of tremor is introduced, which will be referred to as deviation of modulation level, or DML. It is worth noticing that this new tremor attribute is based on the capability of the suggested tremor-estimator to produce time-varying modulation frequency and modulation level, overcoming the limitations of short signal duration.

Fig.6 combines the two characteristics, the modulation level and the DML in one graph for normophonic and dysphonic speakers; each data point has two tremor coordinates; the DML and the mean modulation level. The arrows show the

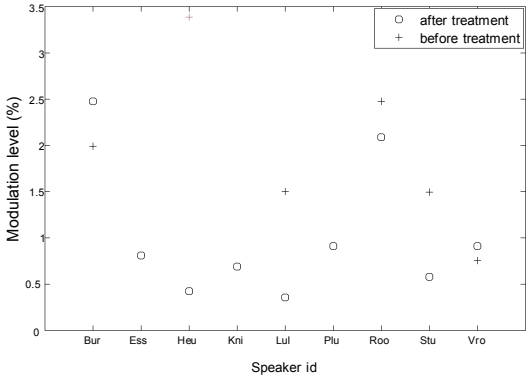


Fig. 5. Modulation level of each dysphonic speaker before and after treatment.

change of the tremor coordinates for dysphonic speakers after treatment. The beginning of the arrow corresponds to the tremor coordinates of the dysphonic speaker before treatment and the end of the arrow to the tremor coordinates after treatment. Each arrow is named after the speaker. The normophonic speakers occupy the low left part of the graph, where the modulation level and the DML take low values, defining therefore a "normophonic area" of these attributes. As it is shown in Fig.6, the untreated dysphonic speakers diverge from the normophonic area. The dysphonic speakers after treatment tend to reach the normophonic region as the arrows show. However, some patients (Roo, Bur) seem to have no improvement. Notice that for some treated speakers there is no estimation of their previous state (before treatment) since, due to the severity of their disease, their phonemes could not be analyzed (small signal duration).

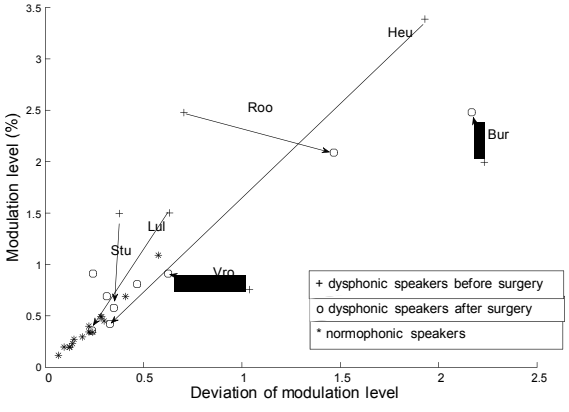


Fig. 6. Mean modulation level as a function of its deviation for normophonic speakers and for speakers with spasmodic dysphonia before and after treatment.

The above analysis suggests that the modulation level and the its corresponding deviation are significant values defining tremor and these attributes can be used either for classifying speakers as normophonic or dysphonic, or for classifying speakers according to the severity of dysphonia. Furthermore,

it may be used as an objective measure for the patient's progress evaluation before and after treatment.

B. Objective Tremor Classification Method

As we saw in the previous section, the tremor signal that modulates an instantaneous component differs significantly in a healthy and in a dysphonic speaker. The outcome of our analysis is that the instantaneous modulation level of the untreated patients with spasmodic dysphonia present high variations in time. A dysphonic speaker appears to have higher modulation level and significant DML than a normophonic speaker. Therefore, we suggest the introduction of a quality indicator that classifies speakers according to their tremor value in their voice. The quality indicator is called Weighted Mean Tremor Value (WMTV) and is defined as:

$$WMTV = w\bar{x} + (1 - w)\sigma(x) \ , \qquad (1)$$

where $\bar{x}$ is the mean modulation level, $\sigma(x)$ the standard deviation of the modulation level of the tremor signal, and w is a weighting factor.

The severity of the spasmodicity of each speaker is ranked using the WMTV with a 40% weighting factor, deriving from the analysis. Our classification is compared with the subjective ranking of tremor for the same speakers and same speech files. Both classifications are presented in Table I. The subjective evaluation was conducted by specialized doctors. In Table I the "-pre" ending corresponds to dysphonic speakers before treatment and the "-pos" to the dysphonic speakers after treatment. For instance, speaker Bur, according to the subjective evaluations, had a slight enhancement after surgery (from 1.00-Burpre to 0.94-Burpos). Notice, that there are differences in the subjective and in the proposed objective classification. However, both evaluations "separate" the patients with severe tremor. For example, both evaluations agree that patients Bur and Roo have high tremor despite treatment and that patients Heu, Stu, Lul, Plu and Ess have low tremor values after treatment. It is found that the correlation between our ranking and the subjective ranking is significant; the correlation coefficient is 0.72 and the p-value is 0.0024.

A) Subjective classification		B) Proposed classification	
	Normalized TR		WMTV
Burpre	1.00	Heupre	1.00
Burpos	0.94	Burpos	0.91
Roopre	0.82	Burpre	0.85
Stupre	0.71	Roopos	0.68
Roopos	0.59	Roopre	0.56
Vropre	0.53	Lulpre	0.39
Vropos	0.47	Vropre	0.37
Heupre	0.41	Stupre	0.33
Knipos	0.41	Vropos	0.30
Lulpre	0.24	Esspos	0.24
Plupos	0.12	Plupos	0.20
Esspos	0.06	Knipos	0.18
Heupos	0.06	Stupos	0.18
Lulpos	0.06	Heupos	0.15
Stupos	0.0	Lulpos	0.11

TABLE I
DYSPHONIC SPEAKERS CLASSIFIC ATION BASED ON: A) SUBJECTIVE EVALUATION, B) DESCENDING WMTV (WEIGHTING FACTOR = 40%)

Fig.7 compares the two evaluations. The ideal match between the two evaluations is the solid line. The closer the markers are to the line the more our method agrees with the subjective evaluations.

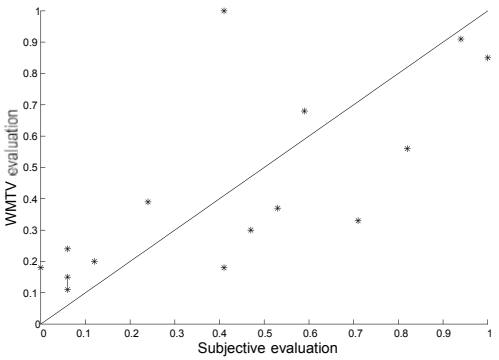


Fig. 7. Subjective evaluation to WMTV evaluation. The solid line corresponds to the ideal match between the two evaluations.

IV. CONCLUSION

Our proposed method aims at estimating tremor in speakers with spasmodic dysphonia. Evaluation results show that it achieves to estimate accurately the time-varying characteristics of tremor. From the analysis in normophonic and dysphonic speakers, a new tremor attribute is introduced, the deviation of the modulation level. This attribute derives from the time-varying characteristics of the modulation level and plays a prominent role in the objective classification of speakers according to their tremor. The two significant attributes, the modulation level and its deviation are combined in one value; the weighted mean tremor value, or WMTV. It was shown that WMTV is a quality indicator of tremor in voice and can be used as an objective measure for evaluating speakers with spasmodic dysphonia.

REFERENCES

[1] P. H. Dejonckere, K. J. Neumann, M.B.J. Moerman, and J.P. Martens. Perceptual and Acoustic Assesment of Adductor Spasmodic Dysphonia Pre-and PostTreatment with Botulinum Toxin. Proceedings Madrid, 2009.

[2] W. S. Winholtz and L. O. Ramig. Vocal tremor analysis with the vocal demodulator. *Journal of Speech Hearing Research*, 35:562–573, 1992.

[3] J. Schoentgen. Stohastic models of jitter. *Journal of Acoustic Society of America*, 109:1631–1650, 2001.

[4] J. Kreiman, B. Gabelman, and B.R. Gerratt. Perception of vocal tremor. *Journal of Speech, Language and Hearing Research*, 46:203–214, 2003.

[5] H. Ackermann and W. Zeigler. Acoustic analysis of vocal instability in cerebellar dysfunctions. *Annals of Otology, Rhinology and Laryngology*, 103:98–104, 1994.

[6] Y. Pantazis, M. Koutsogiannaki, and Y. Stylianou. A Novel Method for the Extraction of Vocal Tremor. In *MAVEBA*, Florence, 2009.

[7] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-FM Signal Decomposition with Application to Speech Analysis. *IEEE Trans. on Audio Speech and Language Processing*, 19(2):290–300, February 2011.

[8] D.I. S Luhring, M. Moerman, J.P. Martens, D. Deuster, F. Muller, and P. Dejonckere. Spasmodic Dysphonia, Perceptual and Acoustic Analysis: Presenting New Diagnostic Tools. *Eur Arch Otorhinolaryngol*, 2009.

[9] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM Estimation for Speech based on a Time-varying Sinusoidal Model. In *Interspeech*, Brighton, 2009.

[10] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

ASSESSMENT OF VOCAL DYSPERIODICITIES IN DISORDERED SPEECH BASED ON EMPIRICAL MODE DECOMPOSITION

A. Kacha[1], F. Grenez[2], J. Schoentgen[2,3]

[1] Laboratoire de Physique de Rayonnement et Applications, Université de Jijell, Jijel, Algeria
[2] LIST Department, Universite´ Libre de Bruxelles, Brussels, Belgium
[3] National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

Abstract: **In this paper, empirical mode decomposition (EMD) is proposed as an alternative method in the framework of acoustic analysis of disordered speech for the purpose of clinical evaluation of voice. The empirical mode decomposition algorithm decomposes adaptively a given signal into oscillation modes extracted from the signal itself. The proposed approach for objective assessment of vocal dysperiodicity consists of two steps. In the first step, the dysperiodicity is estimated by using the generalized variogram. In the second step, the estimated dysperiodicity is decomposed into several narrow-band oscillating components via the EMD algorithm followed by a computation of the segmental signal-to-IMF ratio (SIRSEG) which is used as an acoustic marker for vocal dysperiodicity assessment. The proposed method is evaluated on a corpus comprising 251 normophonic and dysphonic speakers. Results show that the acoustic marker involving some selected IMFs outperforms that obtained from a full-band analysis in terms of correlation with perceptual scores.**

Keywords: **vocal dysperiodicities, empirical mode decomposition, disordered speech.**

I. INTRODUCTION

Objective measures for quality assessment of voice of dysphonic speakers provide a severity index of the disorder which enables clinicians to monitor the progress of patients and document quantitatively the perceived degree of hoarseness. Different acoustic markers have been used to characterize the speech of dysphonic speakers; however the reliability and accuracy are still an issue.

Recent approaches for vocal dysperiodicities esimation in continuous speech are based on long-term prediction [1] and generalized variogram [2]. In [3], the performance of multiband segmental signal-to-dysperiodicity ratio has been investigated in terms of the correlation with scores of perceived hoarseness on a corpus comprising a total of 22 speakers with normophonic and dysphonic subjects.

More recently, the performance of the methods has been evaluated on much larger corpora (a total of over 900 speakers sustaining sounds and producing connected speech). It has been concluded that multi-band segmental signal-to-dysperiodicity ratio correlates more strongly with the perceptual assessment of the degree of hoarseness than the full-band analysis [4].

Although, experimental results obtained in [4] have shown that multiband analysis outperforms one full-band analysis, the multiband approach requires a large amount of data. Indeed, to avoid the risk of overfitting when carrying out multi-band analysis, one should have a large size corpus that enables to compute multiple regression coefficients so that this analysis can not be carried out when only limited data are available. Moreover, the selection of the frequency bands for the analysis is questionable.

In this paper, we propose an alternative approach based on the empirical mode decomposition (EMD) algorithm [5] to filter the dysperiodicity estimated via the generalized variogram. Rather than a priori fixing the number of filters and their corresponding frequency bands, the method decomposes adaptively the dysperiodicity in many narrow-band components, named intrinsic mode functions (IMFs), the number and the frequency content of which are data-driven. A segmental signal-to-IMF ratio (SIRSEG) can be defined for each IMF.

II. METHODS

A. Vocal Dysperiodicity Estimation

Voiced speech is characterized as a quasi-cyclic waveform. When the speech signal is cyclic and the cycle amplitudes change smoothly, it is possible to predict the present cycle on the base of some previous cycle. Most of the disorders originate from the vocal system and frequently result in an increase in the dysperiodicity of voiced speech sounds.

the dysperiodicity may be estimated via the minimum of the following expression named the generalized variogram [2]:

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

$$\hat{\gamma} = \min_{T} \left[\sum_{n=0}^{N-1} \left(x(n) - a x(n-T) \right)^2 \right] \quad (1)$$

with $-T_{\max} \le T \le -T_{\min}$ and $T_{\min} \le T \le T_{\max}$.

The weight a is a positive number to be computed and index n positions speech samples within the analysis frame. Boundaries T_{min} and T_{max} are, in number of samples, the shortest and longest acceptable glottal cycle lengths. They are fixed to 2.5 *ms* and 20 *ms*, respectively (i.e. 50 Hz ≤ F0 ≤ 400 Hz).

B. EMD-based Vocal Dysperiodicities Assessment

The proposed method for objective measure of vocal dysperiodicity in disordered speech consists of two steps. In the fist step, the generalized variogram-based approach is used to estimate the dysperiodicity. In the second step, the dysperiodicity is decomposed adaptively into locally oscillating components called intrinsic mode functions (IMFs) via the empirical mode decomposition (EMD) algorithm developed by Huang et al. for multi-component nonlinear and nonstationary signals analysis [5]. The EMD is a time-frequency analysis tool that does not require a priori fixed function basis like conventional time-frequency representations (e.g. Wigner-Ville distribution or the wavelet transform). The EMD effective tool to decompose the dysperiodicity into several narrow-band components so that each component can be processed separately. Each IMF component has a zero-mean value and only one extreme between zero-crossings.

Let $e(n)$ be the energy-normalized vocal dysperiodcitty estimated via the generalized variogram-based approach. The iterative sifting process for estimating the IMFs involves the following steps:

1. Initialize the algorithm: j=1, initial residue $r_0(n)=e(n)$ and fix the threshold δ

2. Extract local maxima and minima of $r_{j-1}(n)$

3. Compute the upper envelope $U_j(n)$ and lower envelope $L_j(n)$ by cubic spline interpolation of local maxima and minima, respectively

4. Compute the mean envelope

$$m_j(n) = \left(U_j(n) + L_j(n) \right) / 2 \quad (2)$$

5. Compute the jth component $h_j(n)=r_{j-1}(n)-m_j(n)$

6. $h_j(n)$ is treated as $r_j(n)$. Let $h_{j,0}(n)=h_j(n)$ and $m_{j,k}(n)$, k=0, 1,. . ., the mean of the upper envelope and lower envelope of $h_{j,k}(n)$, then compute $h_{j,k}(n)=h_{j,k-1}(n)-m_{j,k-1}(n)$ until

$$SD_k = \sum_{t=0}^{T} \frac{\left| h_{j,k-1}(n) - h_{j,k}(n) \right|^2}{\left(h_{j,k-1}(n) \right)^2} < \delta \quad (3)$$

7. Compute the jth IMF as $\mathrm{IMF}_j(n)=h_{j,k}(n)$

8. Uptade the residue $r_j(n)=r_{j-1}(n)-\mathrm{IMF}_j(n)$

9. Increase the sifting index j and repeat steps 2 to 8 until the number of local extrema in $r_j(n)$ is less than 3.

Each IMF is a narrowband AM-FM component that can be characterized by its instantaneous frequency. The signal can be reconstructed exactly by summing all the J IMFs

$$e(n) = \sum_{j=1}^{J} IMF_j(n) + r_{J+1}(n) \quad (4)$$

To summarize the amount of dysperiodicity within an utterance, for each IMF component, segmental signal-to-IMF ratio (SIRSEG) is computed as the ratio of the signal power to the IMF power:

$$SIRSEG_j = \frac{10}{K} \sum_{k=0}^{K-1} \log \frac{\sum_{n=Mk}^{Mk+M-1} x^2(n)}{\sum_{n=Mk}^{Mk+M-1} IMF_j^2(n)}, \; j = 1, \cdots, J \quad (5)$$

where M is the segment length in samples and K is the number of segments in an utterance.

The acoustic marker SIRSEG provides an objective measure of the relative power of a narrow-band filtered version of dysperuidicity compared to the power of the signal.

C. Corpus and Perceptual Assessment

Speech data used in the present study were used elsewhere [4]. The corpus comprises concatenations of two Dutch sentences followed by vowel [a]. Dutch sentences ("Papa en Marloes staan op het station. Ze wachten op de trein.") have been produced by 28 normophonic and 223 speakers with different degrees of dysphonia. Five judges have evaluated the corpus involving the concatenation of the sentences and vowel [a] perceptually. The five judges are professional voice therapists with at least five years of experience in clinical voice quality ratings. Each judge has rated, from 0 to 3, the item "grade" of the (G)RABS scale. "Grade" represents the degree of hoarseness or voice abnormality. The five perceptual scores per stimulus have been averaged.

III. RESULTS AND DISCUSSION

The performance of the acoustic cue obtained using empirical mode decomposition-based filtering is investigated and compared to that of the segmental

signal-to-dysperiodicity used in [5]. For each stimulus, the dysperiodicity has been estimated via the generalized variogram and the dysperiodicity traces have been decomposed using the EMD algorithm. The decomposition of the dysperiodicity via the EMD algorithm yields more than 20 IMFs, however, only the first ten IMF components have been used in our investigation because they contain more than 90 % of the total energy of the dysperiodicity.

In order to determine the contribution of each of the ten IMFs and to investigate whether summing some specific IMF components enables improving the overall correlation with perceptual scores of hoarseness, a set of 45 traces per stimulus have been formed as follows. For each subset of the first k IMFs (k=1 . . . 10), k traces are obtained by summing the last j IMFs (j=1 . . . k).

For each trace, segmental signal-to-IMF ratio has been computed for the whole data. Pearson's product moment correlations of segmental signal-to-IMF values with average hoarseness scores of the corpus are shown in Fig. 1. The labels of the horizontal axis are the values of the lower order of the IMF included in the sum when forming a trace and the labels of the vertical axis are the values of the correlation. An acoustic marker in a good agreement with the quality of voice must be strongly correlated to scores of the perceived degree of hoarseness. Fig. 1 shows that the correlation tends to increase in absolute value as the IMF components of orders 6 to 8 are included in the trace. The higher correlation R=-0.74 is attained at the IMF order 8 and the correlation decreases beyond this order. The empirical mode decomposition-based filtering results in a higher correlation between SIRSEG values and hoarseness scores for the IMF components 6 to 8 than the one (R=-0.7) obtained by a full-band analysis, i.e., between SDRSEG and hoareness scores and the difference is statistically significant. For more illustration, Fig. 2 shows the correlation of SIRSEG values with average hoarseness scores for each IMF component and Fig. 3 displays the estimated SIRSEG values versus the average hoarseness scores.

The quartiles of SIRSEG values for each IMF component are shown in Fig. 4. As can be observed a given quartile of SIRSEG values takes different values for different IMF components. The quartiles attain their minimum values at the fifth IMF component and tend to increase as the order of the IMF increases or decreases. For IMF components of order lower than 5 the difference between the minimum and the maximum values of SIRSEG is greater than this difference for IMF components of order higher than 5 which is an indication of the difference in the concentration of SIRSEG values, i.e. the difference in the correlation between SIRSEG values and average hoarseness scores for IMF components of order greater than 5 and those of order lower than 5.

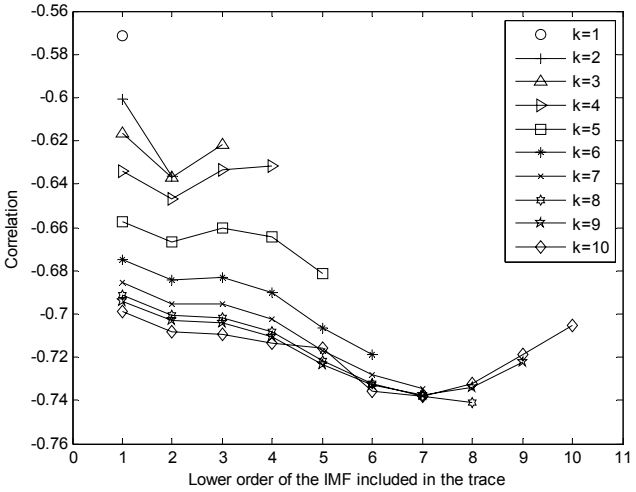


Figure 1: Illustration of the effect of the different IMF components on the value of the correlation between SIRSEG values and average hoarseness scores assigned by five judges. Each symbol corresponds to some subset of IMF components included in the computation of the SIRSEG values.

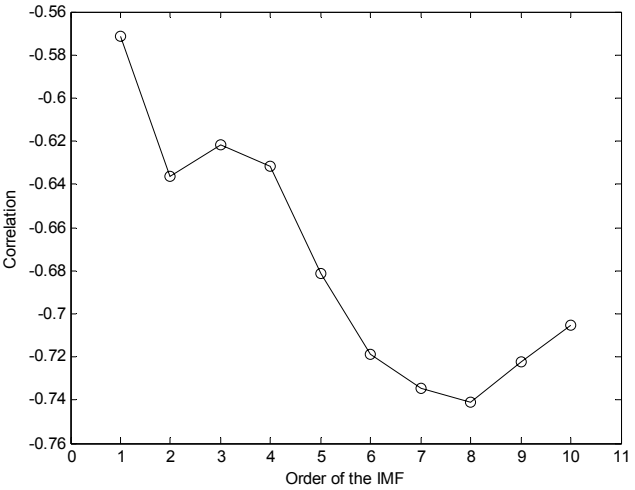


Figure 2: Pearson's product moment correlation between SIRSEG values and average hoarseness scores assigned by five judges for different IMF components. Only one IMF component is included in the computation of the SIRSEG values.

A possible explanation of the strong correlation between SIRSEG values and average hoarseness scores is as follows. The estimated dysperiodicity is still correlated to the signal even this correlation is weak. As a consequence, each IMF obtained from the decomposition of the dysperiodicity contains an amount of the signal which is the smallest for the IMF components of orders 6 to 8 resulting in the highest correlation of SIRSEG values with average hoarseness scores.

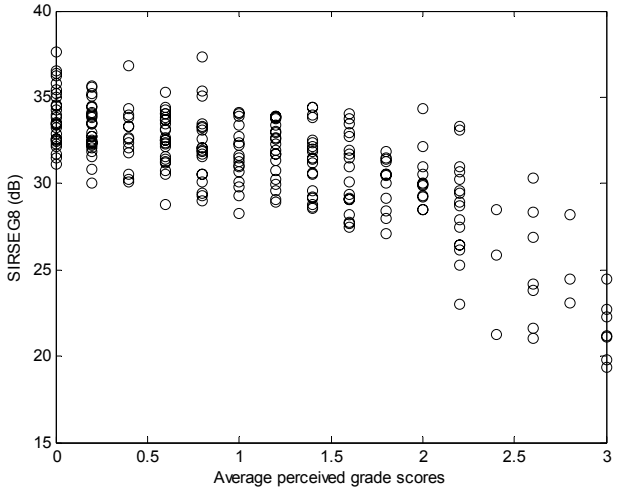


Figure 3: Estimated SIRSEG vs average perceived grade scores for the IMF component *j*=8.

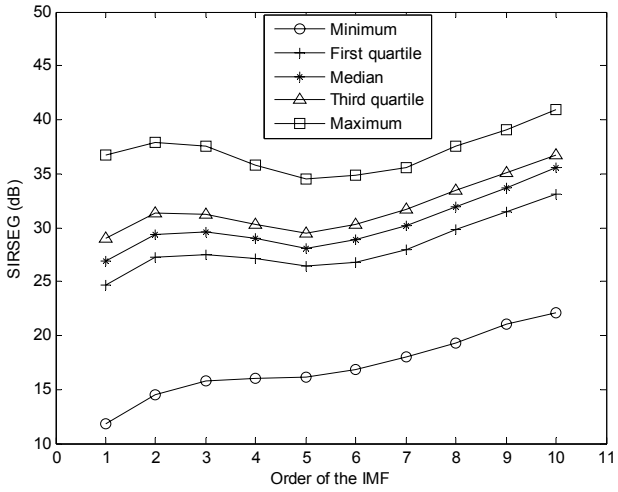


Figure 4: Quartiles of segmental signal-to-IMF values for different IMF components.

In order to investigate the characteristics of the pertinent IMF components in terms of frequency bands, the power spectrum of IMFs 6, 7 and 8 for a speaker that has been assigned an average perceived grade score of 1 is shown in Fig. 5. The central frequencies of the respective bands are 1030 Hz, 690 Hz and 480 Hz. These results are in a good agreement with the values of the cut-off frequency of the filter that gives rise to high correlation between the acoustic marker and the average perceived grade scores [4].

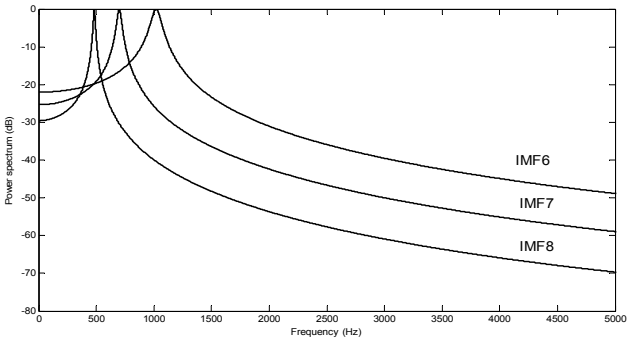


Figure 5: Power spectum density of IMF components 6, 7 and 8 for a speaker that has been assigned an average perceived grade score of 1.

IV. CONCLUSION

In this paper, empirical mode decomposition algorithm has been used to analyse vocal dysperiodicities in disorderd speech. The EMD has been used as a filter bank to decompose the dysperiodicity estimated by means of the generalized variogram into several narrow-band oscillating components (IMFs) and then segmental signal-to-IMF ratio corresponding to each IMF component has been used as an acoustic objective measure for vocal dysperiodicity assessment. The proposed approach has been tested on a large corpus comprising 251 normophonic and dysphonic speakers. Experimental results have shown that for some selected IMFs, EMD-based filtering results in a stronger correlation between SIRSEG and average scores of perceived hoarseness than that achieved by the conventional full-band generalized variogram analysis.

REFERENCES

[1] F. Bettens, F. Grenez, J. Schoentgen, "Estimation of vocal dysperiodicities in connected speech by means of distant-sample bi-directional linear predictive analysis," *J. Acoust. Soc. Amer.*, vol. 117, pp. 328-337, 2005.

[2] A. Kacha, F. Grenez, J. Schoentgen, "Estimation of Dysperiodicities in Disordered Speech," *Speech Communication*, vol. 48, pp. 1365-1378, 2006.

[3] A. Kacha, F. Grenez, J. Schoentgen, "Multiband Frame-Based Acoustic Cues of Vocal Dysperiodicities in Disordered Connected Speech," *Biomedical Signal Processing and Control*, vol. 1, pp. 137-143, 2006.

[4] A. Alpan, Y. Maryn, A. Kacha, F. Grenez, J. Schoentgen, "Multi-band dysperiodicity analyses of disordered connected speech," *Speech Communication*, vol. 53, pp. 131-141, 2011.

[5] Huang N.E. et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. R. Soc. London Ser. A* vol. 454, pp. 903-995, 1998.

[6] Flandrin P., Rilling G., Conçalvès P., "Empirical Mode Decomposition as a Filter Bank," *IEEE Signal Proc. Letters*, vol. 11, pp. 112-114, 2004.

ACOUSTIC ASSESSMENT OF SPASMODIC DYSPHONIA USING A NEW MULTIPURPOSE VOICE ANALYSIS TOOL

A. Giordano[1], P.H. Dejonckere [2,3,4] , C. Manfredi[1]

[1] Department of Electronics and Telecommunications, Università degli Studi di Firenze, Firenze, Italy
2Utrecht University, Utrecht, The Netherlands
3Federal Institute of Occupational Diseases, Brussels, Belgium
4Catholic University of Leuven, Leuven, Belgium

Abstract: **A new multipurpose voice analysis tool named BioVoice2, suited for the analysis of strongly irregular signals and long sentences, is applied on voices of patients diagnosed with adductor spasmodic dysphonia before and after treatment with botulinum toxin injection. The speech material consists of 40 short German sentences phonetically selected to be constantly voiced. Nine acoustic parameters were taken into account from all those estimated with BioVoice2. Significant improvement of voice quality was estimated by a subset of these parameters related to increased voicing, improved regularity of vocal fold vibration, reduction of spasms and faster speech rate. BioVoice2 proves to be a useful tool for objectifying voice quality also in case of strong signal irregularity.**

Keywords: **Spasmodic dysphonia, voice analysis, acoustic parameters, botulinum.**

I. INTRODUCTION

Spasmodic dysphonia (SD) is a particular voice disorder characterized by involuntary movements of one or more muscles of the larynx during speech.
The most common form of this pathology is the adductor SD (ADSD) that is considered in this study. The ADSD is expressed with different severity from case to case, from mild cases that present only a slight tremor of the voice and occasional breaks to cases where severe spasms of the vocal cords make it impossible to speak, preventing airflow through the glottis. In such cases the patient's work and social life are compromised and this can also frequently lead to severe depression [1].
With (AD)SD, deviant acoustic events as aperiodicity, phonatory breaks and frequency shifts perturb fluency and intelligibility. These voices thus require specific acoustic parameters for an exhaustive analysis [2-4].
The present study is based on a new multipurpose voice analysis tool, named BioVoice2, developed under MatLab environment, capable to deal with highly irregular voice signals as those under study.
The aim of the present study is to test the ability of BioVoice2 to evaluate the improvement in patient's voice quality using objective parameters and, accordingly, the effectiveness of the medical treatment that consist of botulinum toxin injection in the vocalic muscles.

II. METHODS

Currently most of the software tools for voice analysis have limitations related to the level of irregularity in the voice and to their applicability to running speech instead of sustained vowels only [5]. To overcome these limitations, we designed a multipurpose program, BioVoice 2 that is applicable to the analysis of a wide range of voice signals, including the analysis of long sentences (several minutes of connected speech) other than short ones or sustained vowels only.
The main targets in designing BioVoice 2 were:

- Implementing a robust and reliable fundamental frequency (F0) estimation and a Voiced/Unvoiced (V/U) selection procedure, applicable to quasi-stationary/noisy signals such as highly hoarse, irregular voices and/or sentences;
- Allowing for the analysis of long sentences (several minutes) other than short ones or sustained vowels only;
- Giving the user a simple Graphic User Interface (GUI) that does not require any manual setting by the user, thus being well suited also for non-expert users.

BioVoice 2 performs the analysis of audio files resulting in objective parameters that are considered useful by clinicians in the diagnosis of voice disorders. It has been successfully tested on synthesized sustained vowels giving better results than most commonly used software tools when applied to strongly irregular and hoarse voice signals [9]. The main parameters of interest in the present work are:

F0: the fundamental frequency is estimated with a two steps procedure. First, Simple Inverse Filter Tracking is performed, obtaining a raw F0 estimation and its range of variation (F_L, F_H) where F_L= lowest F0 value and F_H =

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

highest F0 value. In the second step, F0 is estimated inside (F_L, F_H) with the Average Magnitude Difference Function (AMDF) approach [9]. The program provides F0 tracking and its mean, standard deviation, minimum and maximum values.

PVF: the ratio between the number of voiced frames and the total number of frames, that depends on the breaks which are present in the voice: the more the pauses, the less the PVF.

PVS: the percentage of voiced speech frames, which means the ratio of voiced frames over the frames that have been classified as speech in a previous step of analysis. Speech frames are those in which the zero crossing rate is less than 3000 zero-crossings per milliseconds and the energy exceeds a threshold value that depends on the signal characteristics. Hence PVS should be higher or equal to PVF. On a sustained vowel and for a healthy voice, PVS is ideally 100%. As a general rule the better the voice, the higher both PVF and PVS.

PFU: the percentage of frames that have an unreliable F0 among the total number of frames. This parameter is therefore a measure of the fundamental frequency F0 instability. Frequency variations make F0 unstable. In a frame F0 is evaluated as unreliable if it has a deviation of more than 25% compared to the average F0 value over all voiced frames. In this case the better the voice, the lower the PFU percentage.

VL90: the 90th percentile of voicing length distribution, defined as the maximum number of consecutive voiced frames found. The sharp breaks featuring the voice of patients with SD reduce this parameter.

Duration: the total time required to the patient for pronouncing sentences. As a general rule a healthy voice, that is more fluent, will have a shorter duration than a pathological one.

In addition BioVoice 2 evaluates the time duration of the voiced and unvoiced part of the signal, and the average length of voiced frames (mean duration of voicing, **MDV**).

Jitter: a measure of the degree of variability of the period length. It gives a measure of the aperiodicity of the signal measuring the changes in fundamental period T0=1/F0 from period to period. Of course, good voices have low jitter. Jitter J is evaluated here according to Eq.1:

$$J = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \qquad (1)$$

Where N is the number of frames and T_i is the i-th period length.

Corrected jitter: the correction means that only frames with reliable F0 are taken in account. F0 is reliable if it has less than 25% deviance from the mean value of F0 of all voiced frames. The formula for Corrected jitter is the same as for the jitter.

NNE: Normalized Noise Energy is a noise estimation method that relies on a comb filtering approach: it is the ratio of the energy between the harmonics and the whole signal energy [8].

Moreover, BioVoice2 allows for the estimation of the signal spectrogram, formants, Power Spectral Density and other parameters related to the kind of voice signal under analysis (adult male, adult female, newborn cry and singing voice) that are not described here. Plots and tables can be displayed, printed and saved in an easy way. Details can be found in [5].

To test the capability of BioVoice2 of analyzing long sentences in a reliable way, 24 audio files (12 pre- and 12 post-treatment) from 12 German patients diagnosed with ADSD are considered here. Each patient read a standardized list of 40 German sentences for a total duration of about 2'30". These sentences are phonetically selected by clinicians for being constantly voiced. This is in fact supposed to increase the sensitivity for detecting interruptions of vocal fold vibrations induced by SD.

Audio files are provided in uncompressed audio wave format with sampling frequency Fs= 44.100 Hz and 16 bit of resolution. All the recordings were made in a quiet room by one of the authors of this work.

III. EXPERIMENTAL RESULTS

Table 1 reports the mean value of the parameters previously described, obtained from pre and post-treatment recordings. From Table 1 a clear trend towards better voice quality is shown (post-treatment values higher or lower than pre-treatment ones, according to the specific parameter).

Table 1 – Mean value of the acoustic parameter computed by BioVoice2.

Parameter		PRE	POST
PVF %	Mean	51.80	69.45
PVS %	Mean	52.82	69.54
PFU %	Mean	50.29	44.81
Jitter %	Mean	14.63	13.76
Corrected Jitter %	Mean	6.12	6.24
VL90 [s]	Mean	0.0008	0.0160
Duration [s]	Mean	142.07	137.68
MDV [s]	Mean	0.277	4.03
NNE [dB]	Mean	-17.49	-17.58
Mean F0 [Hz]	Mean	180.65	188.62
Std F0 [Hz]	Mean	46.05	50.05

These results show that the parameters PVF, PVS, VL90 as well as MDV are strongly indicative of voice improvement, while for the other parameters the pre-post difference seems less significant.

Figure 1 shows the difference between pre- and post-treatment values of the most relevant acoustic parameters that are PVF, PVS, VL90 and MDV.

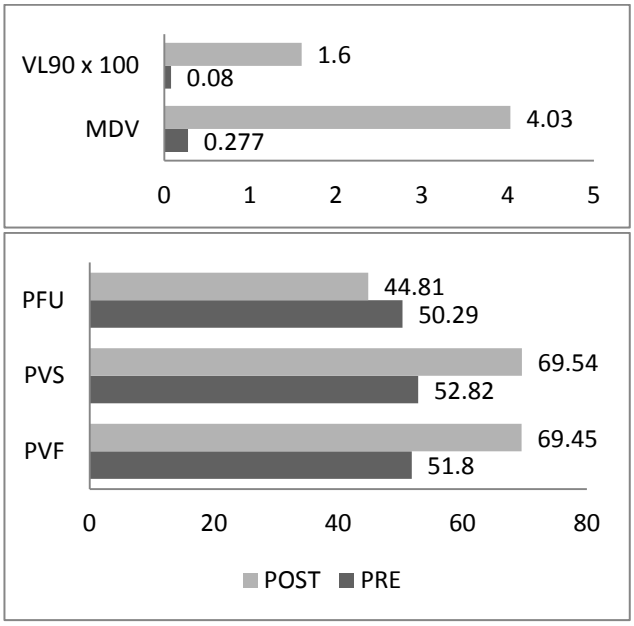


Fig. 1: Upper: mean value of VL90 and MDV (seconds). Lower: mean value of PVF, PVS, PFU (%) for pre-post treatment data.

Moreover the Wilcoxon test was applied on each parameter separately. Results are reported in Table 2.

Table 2 – Results of the Wilcoxon test for all the acoustic parameters.

Wilcoxon's test	
Parameter	***P***
meanF0	0.5186
Std F0	0.3804
PVF	0.0269
PVS	0.0161
PFU	0.3394
Jitter	0.8501
Corrected jitter	0.8984
NNE	0.9697
VL90	0.0342
Duration	0.7334
MDV	0.0425

As expected, only few of the acoustic parameters reveal a significant post- vs. pre- improvement. Specifically these parameters are: PVF, PVS, VL90 and MDV. In particular Jitter, Corrected jitter, PFU, NNE and even duration show no statistically significant differences and are thus not suited for evaluating the improvement of voice quality with the present data. As these parameters have different measurements units and ranges a standardization step was performed according to the following equation [7]:

$$Z_i = \frac{x_i - \bar{x}}{\sigma} \quad (2)$$

Where x_i is the variable to be standardized, $\bar{x}$ is its mean value and σ is its standard deviation.

Figure 2 shows the boxplot of pre- post-treatment data for all the acoustic parameters considered here. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles respectively, and the whiskers extend to the most extreme data points that are not considered outliers. The small circles are the outliers. Data are standardized according to Eq.2.

The plot confirms the best results obtained with parameters PVF, PVS, VL90 and MDV.

IV. Discussion

Results in Table 2 show that only four of the whole acoustic parameters considered here are capable to point out a significant post- vs. pre-treatment improvement in voice quality. These parameters are all related to the increased voicing capability of the patient after medical treatment.

Hence only the acoustic parameters that are in some way related to the selection of voiced/unvoiced parts of the signal are successful in the analysis of a long sentence, while other, and also jitter, seem to have less relevance. This result suggests that a different analysis should be performed on fluent speech other than that usually made on sustained vowels or short sentences.

Results are in agreement with previous studies made on the same speech material where a different analysis program was used [7]. However, differently from [7], with BioVoice2 more parameters, such as a noise measure, F0 and its standard deviation can be included in the analysis.

Moreover, a new parameter was introduced here for the first time, namely the mean duration of voiced frames, MDV. From the preliminary results presented here (Table 1 and Table 2), this parameter seems indeed to be very promising in evaluating the quality of voice in long sentences.

V. Conclusion

A new multipurpose voice analysis tool is presented here and its performance is evaluated on fluent speech coming from patients affected by adductor spasmodic dysphonia. Even if the data set consists of a limited number of patients, significant changes in the value of acoustic parameters were found comparing the pre- and post-treatment recordings, pointing out the improvement of voice quality after botulinum toxin treatment. Some parameters such as jitter, already proved valid in the analysis of short sentence or sustained vowels, seem to lose meaningfulness when evaluated on long sentences. Even the PFU parameter, that is a measure of the

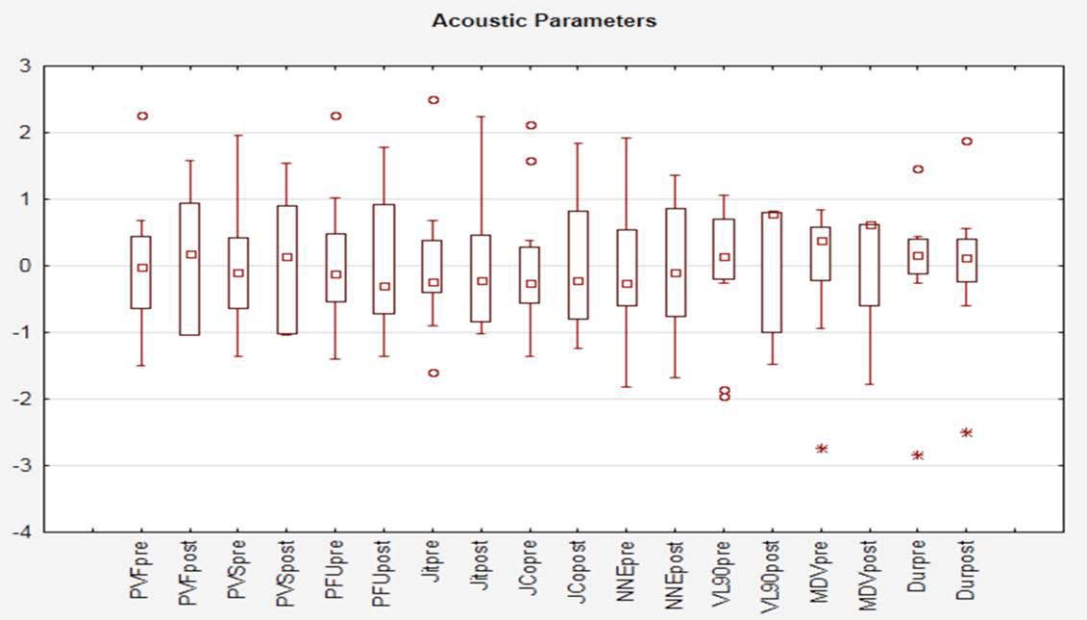


Fig. 2: Boxplot of pre-post treatment acoustic parameters.

fundamental frequency instability, seem to lose its capability in evaluating the voice signal in long sentences.

However, the proposed tool is successful in objectifying the increased voicing and the improved regularity of vocal fold vibration after treatment. One newly defined parameter, the mean value of voiced frame duration, seems very promising in evaluating voice quality improvement when applied to long sentences. Future work will be devoted to refining the tool in order to reduce the computational time while preserving its high resolution capabilities and robustness against noise. The tool will be also tested on a new corpus of synthetic signals with varying F0 and formants that should mimic fluent speech.

REFERENCES

[1] Baylor CR, Yorkston KM, Eadie TL. *"The consequences of spasmodic dysphonia on communication related quality of life: A qualitative study of the insider's experiences."* J. Comm. Disorders. 2005; 38:395–419.

[2] Dejonckere P.H., Bradley P., Clemente P., Cornut G., Crevier-Buchman L, Friedrich G, Van De Heyning P, Remacle M., Woisard V.,*"A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessments techniques"*, Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS), *Eur. Arch. Otorhinolaryngol.* 258, pp.77-82, 2001.

[3] Dejonckere P.H., *"Critères acoustiques de fluence pour l'évaluation des dysphonies spasmodiques. In : Voix parlée et chantée. "* C. Klein – Dallant, Ed. Paris. 63 – 73. ISBN 978-2-9528061, 2007.

[4] Sapienza CM, Cannito MP, Murry T, Branski R, Woodson G : "*Acoustic variations in reading produced by speakers with spasmodic dysphonia pre-Botox injection and within early stages of post-Botox injection*." J Speech Language Hearing Res 45: 830 – 843, 2002.

[5] Manfredi C, Giordano A, Schoentgen J, Fraj S,Bocchi L, Dejonckere PH, "*Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools*" Biomedical Signal Processing and Control, 2011 (in print).

[6] Siemons-Luhring DI, Moerman M, Martens JP, Deuster D, Muller F, Dejonckere PH, *"Spasmodic dysphonia, perceptual and acoustic analysis: presenting new diagnostic tools"* European Archives of Oto-Rhino-Laryngology, Vol. 266, n. 12, 2009.

[7] Dejonckere PH, Moermann KJ, Merman MBJ, Martens JP, *"Perceptual and acoustic assessment of adductor spasmodic dysphonia pre- and post-treatment with botulinum toxin"*, Proc. 3rd AVFA International Worshop, 18th-20th May 2009, Madrid (Spain).

[8] Kasuya H, Ogawa S, Mashima K, Ebihara S, *"Normalised Noise Energy as an Acoustic Measure to Evaluate Pathologic Voice"*, J. Acoust. Soc. Am., vol. 80, n.5, p.1329-1334, 1986.

[9] Manfredi C, Giordano A, Schoentgen J, Fraj S, Bocchi L, Dejonckere PH *"Reliability of voice analysis software tools for highly irregular signals Part II: the effect of noise"* Logopedics Phoniatrics Vocology, 2011 (in print).

Session V:
Professional voice

ACOUSTIC ANALYSIS OF VOCAL FATIGUE IN PROFESSIONAL VOICE USERS

K. V. Evgrafova[1], V. V. Evdokimova [2]

[1]Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia
[2]Department of Phonetics, Saint-Petersburg State University, Saint-Petersburg, Russia

Abstract: **Vocal fatigue is a voice symptom which is frequently reported by professional voice users. Teachers, singers, actors and other professions that require prolonged voice use are especially at-risk group. The vocal fatigue results in auditory perceptual and acoustic changes in the voice signal and can lead to serious pathological conditions. The present study has examined acoustic manifestations of the vocal fatigue in pronunciation teachers who seem to be particularly susceptible vocal and articulatory fatigue. In the paper detailed acoustic analysis of the data obtained is presented. The results of the acoustic analysis showed a consistent dependency between acoustic parameters and vocal fatigue.**

Keywords: vocal fatigue, acoustic analysis, professional voice users

I. INTRODUCTION

Voice problems are known to be common among professional voice users worldwide. Teachers, singers, actors and other professions that require prolonged voice use are especially identified as an at-risk group for developing vocal disorders. A voice symptom which is frequently reported by professional voice users is vocal fatigue which is a complex multifaceted phenomenon that presents a challenge for both research and clinical practice. The symptoms of vocal fatigue are various and explained by the physiologic mechanisms of vocal production. There exist many studies on vocal fatigue providing various concepts of the phenomenon. However, they do not offer a universally accepted definition of it. The vocal fatigue can be viewed either as a voice disorder caused by other pathological voice conditions or as a separate voice problem resulting from prolonged and excessive voice use [10]. In this study the vocal fatigue is understood as a separate phenomenon caused by excessive professional voice load which results in auditory perceptual and acoustic changes in the voice signal and can lead to serious pathological conditions.

Teachers form a large group of voice professionals and their voice problems have been focused on in many studies [3-5, 7, 8]. The present study has examined acoustic manifestations of the vocal fatigue in phoneticians who teach pronunciation. The pronunciation teachers seem to be particularly susceptible to fatigue. They have to repeat articulation drills in front of the students many times and correct continuously their pronunciation which demands a high level of vocal effort and excessive muscular tension of articulators. As a consequence of this vocal overloading the pronunciation teachers often suffer from dysphonia and benign lesions such as nodules. Identifying vocal fatigue in its initial stage is important to prevent voice disorders. Consequently, objective methods to evaluate voice quality under fatigue are required. Acoustic measures could be used as objective criteria to identify at-risk professionals and facilitate intervention strategies to prevent pathological conditions.

II. METHODS

A. Subjects

The methodologies that attempt to induce vocal fatigue in experiment participants vary across numerous works on the vocal fatigue [1-9]. In most studies the vocal fatigue is induced artificially as a result of reading or speaking tasks of various types. The results described are inconsistent and often conflicting.

The conditions of our experiment seem to be more realistically challenging.

10 female teachers were recorded before and after their workdays. All the participants were pronunciation teachers at the Department of Phonetics, Saint-Petersburg State University with average work experience of 7 years. No one had pathological voice problems.

B. Protocol

The participants were asked to read at habitual loudness a four minute phonetically representative text before classes in the morning. After continuous teaching for 8 hours they were asked to record the same text. Each of the participants reported symptoms of a high degree of vocal fatigue after the workday such as a high level of muscular tension/discomfort, hoarse voice quality, breathy voice quality, unsteady voice, inability to maintain typical pitch, dry throat etc. The recordings were made in the recording studio at the Department of Phonetics, Saint-Petersburg State University. Multi-channel recording system Motu Traveler, capacitor

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

microphone AKG and WaveLab program were used. The recordings have a sample rate of 44100 Hz and a bitrate of 16 bits.

C. Material Annotation

To perform the detailed analysis of the non-fatigued and fatigued speech the recorded material was annotated at 6 levels. The annotation captured the maximum amount of phonetically and prosodically relevant data. The six annotation layers are as follows:

Layer 1 – pitch marks;

Layer 2 – phonetic events labeling;

Layer 3 – real phonetic transcription (it is performed manually and reflects the sounds actually pronounced by the speakers);

Layer 4 – ideal phonetic transcription (this layer is automatically generated by a linguistic transcriber in accordance with a canonical set of rules);

Layer 5 - orthographic transcription;

Layer 6 – prosodic transcription.

Layers 1 and 2 contain information on various phonetic events: epenthetic vowels, laryngalization, and glottalization etc. The phonetic events were annotated manually by expert phoneticians.

Prosodic transcription on Layer 6 includes labels for pause and tone unit boundaries and labels for non-speech events such as breathing, cough etc.

The fundamental frequency periods were detected automatically by means of the Wave Assistant program. The results of the automatic procedure were checked and corrected manually.

Layer 3 contains narrow phonetic transcription. It reflects the sounds actually pronounced by the subjects. The 'ideal' transcription found at Layer 4 was generated in accordance with a set of phonological rules without reference to the actual sound. As a result, Layer 4 contains a canonical phonetic transcription of the speech sample. The transcription symbols used were a version of SAMPA for the Russian language. Symbols for vowels contained indication of the sound's position regarding stress. To produce the real phonetic transcription, the speech signal was manually segmented, transcribed and peer-revised by expert phoneticians.

Ideal phonetic transcription was generated automatically by an automatic transcriber. The labels were placed automatically to coincide with the label positions produced manually on the real transcription layer. Procedure of automatic labeling is based on calculating the Levenshtein distance. Automatic labeling is not perfect due to the mismatch of ideal and real phonetic transcriptions and drawbacks of the automatic transcriber. Therefore, the results of the automatic procedure were further manually corrected.

Fig. 1 below shows the multilayered annotation of the sound material.

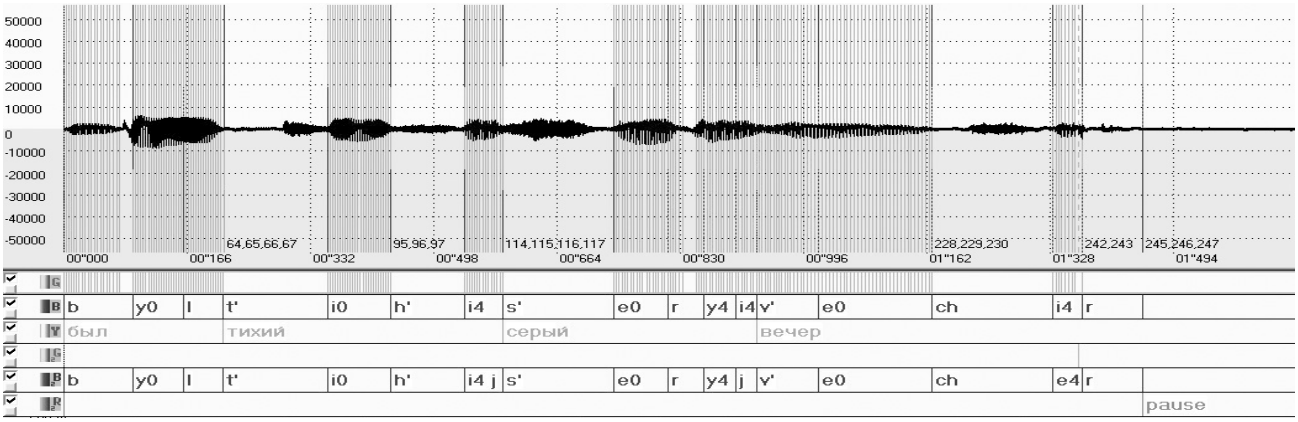


Figure1. The Multilayered Material Annotaion.

III. RESULTS

The vowels and sonants in the before (*non-fatigued* voice) and after (*fatigued* voice) recordings were analyzed for mean fundamental frequency (F0), mean duration (vowels, consonants, pauses), intensity range, pitch range values. Variables of maximum and minimum F0 were also obtained. The vowels were additionally analyzed for perturbation (jitter and shimmer). The acoustic features which correlated with vocal fatigue state were extracted from the recorded material.

A. Fundamental Frequency Variation (non-fatigued/fatigued speech)

The analysis of F0 features shows that the mean pitch value tends to be higher in fatigued speech across all the subjects. The pitch range increases significantly due to the increase of upper range value. The mean lower range value stays practically unchanged.

Table 1 shows the mean fundamental frequency variation across all the subjects in both material types.

Table 1. The Mean Fundamental Frequency Variation

Type of Material	Mean pitch, Hz	Pitch range, Hz	Pitch max, Hz	Pitch min, Hz
Non-fatigued	242	212	351	139
Fatigued	253	308	445	137

Among common fatigue symptoms which are frequently reported by researchers there is a creaky voice quality which is marked by significant decrease in pitch value and pitch breaks (laryngalization) [6,7,10].

However, the analysis of our material shows that the mean duration of laryngalized speech segments turns to be less in fatigued speech than that of in non-fatigued one.

Table 2 shows the ratio of laryngalized speech segments to the whole text recorded.

Table 2. The ratio of laryngalized speech segments to the whole text recorded

Type of Material	Laryngalization, %
Non-fatigued	**1,5**
Fatigued	**1,2**

B. Duration Variation (non-fatigued/fatigued speech)

The tables below show the mean variation of sound duration in the fatigued speech in comparison with that of the same sounds in the non-fatigued speech.

Table 3. The Variation of Vowel Duration (fatigued speech)

Vowel	Duration Increase		Duration Decrease	
	%	ms	%	ms
a	3	2,4		
e	10.2	6,2		
i	7.1	3,6		
o	14.7	11		
u			1.3	0,4
ɨ	6.1	3,1		
All vowels	6	4,3		

As it is shown in the table above, all vowels increased in duration except /u/. The increase ranges from 3% to 10.2%.

The duration of consonants tends to increase as well. Table 4 shows the duration variation across different types of consonants according to the manner of production. The duration did not vary with voiced/voiceless sound quality.

Table 4. The Increase of Consonant Duration (fatigued speech)

Consonant Types	Duration Increase	
	%,	**ms**
Stops	10,3	7,5
Fricatives	10,7	9,3
Affricates	13,7	4,8
All consonants	7,4	5,2

To obtain more data on duration variation in the fatigued speech, pauses were also analyzed. The number of pauses in the recorded material varied with a subject. Some of them made more pauses in the fatigued state (in comparison with a non-fatigued one), while the others made fewer pauses when fatigued. However, in both cases the mean pause duration increased in the fatigued speech. That means that under fatigue it took a subject more time to pause.

Table 5. The Mean Pause Duration (both types of material)

Type of Material	Mean Pause Duration, ms
Non-fatigued	478
Fatigued	567

C. Articulatory Fatigue Manifestations

The comparison of ideal and real phonetic transcription showed the following changes of articulation caused by fatigue.

In table 6 we compare the ideal phonetic transcription reflecting the way the speech sample is supposed to be pronounced according to the canonical transcription rules of the Russian language and the real phonetic transcription reflecting the way it actually was pronounced by the subjects recorded.

Table 6. Ideal vs. Real Transcription.

	Total	Match	Mismatch	Elision
Non-Fatigued	100	84.7	9.05	6.25
Fatigued	100	81.2	10.8	8

Table 6 reveals that percentage of phoneme mismatch (the number of the expected sounds replaced by other sounds) and elisions (the number of the expected sounds which are actually not pronounced at all) is higher in the state of fatigue.

IV. Discussion

The voice acoustic analysis performed before and after the working day can contribute to objective voice examinations useful in diagnosis of dysphonia among the

pronunciation teachers. The pronunciation teachers seem to be a most highly vocally demanding profession and the fact should be taken into account in developing safety work standards and regulations. The 8 hour work load a day is obviously excessive and can lead to pathological conditions.

The perspectives for future works are 1) to test more subjects including male ones, 2) to investigate degrees of vocal fatigue and correlating acoustic parameters, 3) to identify critical threshold of vocal fatigue basing on acoustic analysis, 4) to investigate whether physiologic and/or neurologic fatigue (e.g. induced by sleep deprivation, physical exercise etc.) causes the same effects on the acoustic signal, 5) to compare the acoustic manifestations of vocal and non-vocal fatigue.

V. Conclusion

The results of the acoustic analysis showed a consistent dependency between acoustic parameters and vocal fatigue. After a working day F0 values were higher, the duration of vowels and consonants increased; pitch and loudness range values increased. Measuring jitter and shimmer did not give consistent results. The differences in the acoustic parameters after a vocally loading working day mainly seem to reflect increased muscle activity as a consequence of excessive vocal loading

As well as acoustic manifestations of the fatigue state it is also necessary to consider articulatory changes as a fatigue symptom. It especially matters for pronunciation teacher profession which demands not only a high level of vocal effort, but also excessive muscular tension of articulators.

References

[1] V. J. Boucher, “Acoustic Correlates of Fatigue in Laryngeal Muscles: Findings for a Criterion-Based Prevention of Acquired Voice Pathologies”, in *Journal of Speech, Language, and Hearing Research,* vol. 51, pp. 1161–1170. October 2008.

[2] M.J. Caraty and C. Montacié, “Multivariate Analysis of Vocal Fatigue in Continuous Reading, *The Proceedings of Interspeech 2010*, pp. 470-473.

[3] B.E. Kostyk and A.P. Rochet, “Laryngeal airway resistance in teachers with vocal fatigue: a preliminary study”, in *Journal of Voice*, 1998, vol. 12, pp. 287–299.

[4] E. Sala, E. Airo, P. Olkinuora et al, “Vocal Loading among Day Care Center Teachers”, in *Logoped Phoniatr Vocol*, 2002,vol. 27, pp. 21–28.

[5] B. Schneider, “Effects of Vocal Constitution and Autonomic Stress-Related Reactivity on Vocal Endurance in Female Student Teachers”, in *Journal of Voice*, 2006, vol. 20, No. 2, pp. 242–250.

[6] R.C. Scherer, I.R. Titze et al, “Vocal fatigue in a professional voice user”, in “Transcripts of the Fourteenth Symposium: Care of the Professional Voice”, New York: The Voice Foundation, 1986, pp.124–130.

[7] R.C. Scherer, I.R. Titze et al, “Vocal fatigue in a trained and an untrained voice user”, in *Laryngeal Function in Phonation and Respiration*, San Diego, Singular Publishing Group, 1991, pp. 533–555.

[8] I.Titze, J.Lemke and D. Montequin, “Populations in the U.S. workforce who rely on voice as a primary tool of trade: a preliminary report”, in *Journal of Voice*, vol. 11, 1997, pp. 254–259.

[9] A.M. Laukkanen, “On speaking voice exercises [academic dissertation]”, in *Acta Universitatis Tamperensis*, ser A, vol. 445, Tampere: University of Tampere, 1995.

[10] N.V. Welham et al, “Vocal Fatigue: Current Knowledge and Future Directions”, in *Journal of Voice*, vol. 17, No. 1, 2003, pp. 21–30.

REAL-TIME EMBEDDED TRACKING OF PATIENT REPORTED VOCAL DISCOMFORT IN PROFESSIONAL SETTINGS

I. Verduyckt [1, 2], C. Rungassamy [1], M. Remacle [2,3], T. Dubuisson [4]
[1]Faculté de psychologie, Université de Louvain, Belgique
[2]Service ORL, Université de Louvain, Cliniques Universitaires de Mont-Godinne, Yvoir, Belgique
[3] Centre d'Audiophonologie, Université de Louvain, Cliniques Universitaires de Saint-Luc, Bruxelles, Belgique
[4] Laboratoire de Théorie des Circuits et Traitement du Signal, Faculté polytechnique, Université de Mons, Belgique

Abstract: **The aim of the present study was to evaluate patient-report of vocal discomfort by means of a portable device, designed for the continuous assessment of voice disorders with real-time coupling of acoustic and patient self-evaluation measures.**
10 teachers were equipped with the portable device embedding our vocal discomfort software during 3 days in their professional settings. They had to note their vocal discomfort during the day on a visual analogue scale (VAS) ranging from 0-100 units, either spontaneously, or following an auditory prompt.
The adequacy of the device and of the software was evaluated by a questionnaire addressing the *wearability* of the device, the *easiness of the software*, the a*dequacy of the scale* and the *subjects' annotation behavior*. The adequacy of the scale was further examined by the analysis of the vocal *discomfort ratings* and their *change in value across time*.
The results show good wearability, easiness, and annotation behavior scores, subjects made regular annotations even without auditory prompting. The discomfort scores generally increased during a working day.
The real-time embedded tracking of patient reported vocal discomfort in professional settings can thus be advantageously performed by a portable device, embedding our auto-evaluation software.
Keywords : **Vocal discomfort, real time embedded tracking.**

I. INTRODUCTION

Voice assessment is usually carried out in voice laboratories. Although it is advantageous because of the possibility to perform measures in a reproducible setting, the assessment is however limited for patients whose voice problems arise only in specific situations, as for example teachers in a working environment [1, 2]. The possibility to complete voice laboratory measurements with real-life assessments would be valuable in the diagnostic phase of a voice disorder, for treatment outcome evaluation and for patient monitoring purposes [3].

We are in the development phase of a portable device, designed for the continuous assessment of voice disorders with real-time coupling of acoustic and patient self-evaluation measures. The aim of the present study was to evaluate the adequacy of this device and the software developed on this platform for patient-report of vocal discomfort.

II. METHODS

Subjects were 10 teachers (8 women, 2 men), mean age 35 (Standard deviation - SD: 8,45). Two were teaching in kindergarten, four in primary schools and 4 in secondary school. All subjects judged their professional voice use as intense, none reported suffering from dysphonia. Each subject was equipped with the portable device, embedding our software allowing the notation of vocal discomfort. The notation is performed by the displacement of a cursor along a VAS ranging from 0 – 100 units divided in three colored compartments labeled "low", "moderate" and "high". A validation button has to be pressed to confirm the notation. The last two notations made by the subject were kept visible on the screen. The position of the cursor and the time in seconds is recorded continuously; every activation of the validation button is registered.

Subjects were tested in their professional settings for three consecutive weeks, always on the same day (eg : one subject was tested on three consecutive Mondays while another was tested on three consecutive Tuesdays). A condition where the subjects were asked to make their vocal discomfort notations spontaneously and a condition where they were reminded every 30 minutes by an auditory prompt were tested in a crossover design where subjects were randomly assigned to either group A (auditory prompt on the 1st 3rd week) or group B (auditory prompt on the 2nd week).

As the final objective of this project is to couple continuous audio-recordings with the auto-evaluation of the patient, a microphone was fixed on the subjects' collar in order to test the entire device, although no

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

sound was recorded at this time of the study. Written and oral information on the use of the device and the software were given to each subject.

The adequacy of the device and of the software was evaluated by a questionnaire addressing the *wearability* of the device, the *easiness of the software*, the a*dequacy of the scale* and the *subjects' annotation behavior*. The questionnaire was answered each test day, answers were given on a 10 cm long VAS. The adequacy of the scale was further examined by the analysis of the *vocal discomfort ratings* and their *change in value across time*.

Moreover, subjects were asked to give us a duty roster for each of the test days where they also could report any comments that could be of interest regarding their vocal use.

III. Results

A. *General results.*

All subjects were able to participate on the three days. 29 out of 30 questionnaires were returned. Discomfort notations were collected on 23 out of 30 days. 6 subjects exited the software by mistake during one or two test days. Subjects wore the device for a mean of 7,3 h (SD: 2h) and made a mean of 10,4 annotations per day (SD 8,6), the mean interval between annotations was 49,4 min (SD: 28,1 min). Validations of the same vocal discomfort value that were made in an interval of less than 10 minutes were not taken into account, indeed subjects 2, 5 and 6 made abnormal amounts of validations in a short duration of time (up to 16 validations in the lap of 123 seconds), which was regarded as an artifact.

The auditory prompts were heard on 4 out of 15 days, and on 2 of those days, the subjects had exited the software by mistake. No computations regarding the prompt condition have thus been carried out.

B. Adequacy of the device and the auto-evaluation software.

The questions regarding the device's *wearability* obtained a mean score of 6,8 each (SD: 3,5) (see Fig. 1).

The questions regarding the easiness *of the software* obtained mean scores of 8 (SD: 2,6), 7,1 (SD: 3,1) and 7,8 (SD: 3) (see Fig. 2).

The questions regarding the *adequacy of the scale* obtained mean scores of 7,9 (SD: 2,7), and 8,1(SD: 2,6) (see Fig. 3).

The questions addressing the *subjects' annotation behavior* obtained a mean score of 5,6 (SD: 2,4) and 6,9 (SD: 3,1) (see Fig. 4).

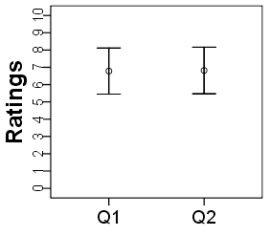


Fig. 1 Wearability (95% confidence interval. *Q1: Was the device bulky to wear? 0: Very bulky, 10 Not bulky at all. Q2: Was the device annoying to wear? 0: Very annoying, 10: Not annoying at all.)*

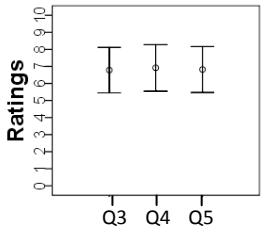


Fig. 2 Easiness of the software (95% confidence interval. *Q3: Was the software easy to understand, Q4: was the cursor easy to move? 0: Not at all, 10: Very easy. Q5: Did you remember to validate after your notation? 0: Never, 10: Always.)*

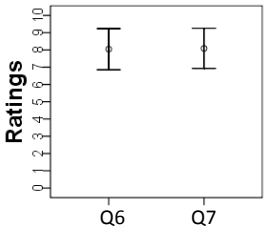


Fig. 3 Adequacy of the scale (95% confidence interval. *Q6: Was the scale adequate for noting your vocal discomfort? Q7: Were the labels helpful in noting your vocal discomfort? 0: Not at all, 10: Very)*

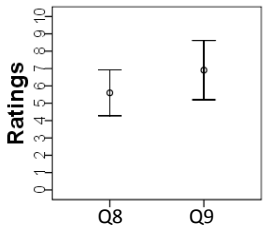


Fig. 4 Subjects annotation behavior (95% confidence interval) *(Q8: Did you forget to note your discomfort? 0: Always, 10: Never. Q9: Did you omit to note your discomfort to avoid getting disturbed in your activities? 0: Always, 10: Never.)*

B. Vocal discomfort measures.

Fig. 5 shows the mean discomfort value (computed on the three consecutive days) for each subject. On the first annotation of the day, the mean discomfort value over the subjects was 11,3 (SD: 10,5), and 41,1 (SD:

30,1) on the last annotation of the day. Subjects two, six and eight did not show an increase of their mean vocal discomfort value during the day.

Looking more closely at their answers (Fig. 6), we see that subject six and eight show overall null and flat vocal discomfort responses. Subject two has a similar answer pattern on day three but on day two, we see that there are increases and decreases of his vocal discomfort until the last hour of the day where it decreases below its initial value; this subject revealed having had a great vocal use on that day, apart from the last hour, where he kept quiet while his students made exercises.

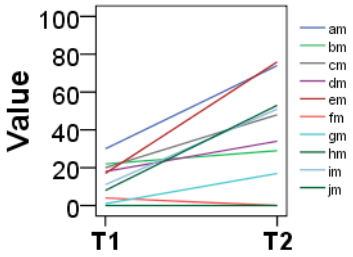


Fig. 5 Mean vocal discomfort values.
(T1: First validation, T2: Last validation.)

Different patterns of vocal discomfort value changes during the day can be observed (Fig. 6), some subjects give flat responses that do not evolve during the day (subjects 8, 6 and subjects 2 and 7 on day 2), some subjects have responses that evolve in a saw tooth pattern (subject two on day three, subject seven on day one and three and subject two on day two and three) and some subjects have responses that evolves gradually over the day (subjects 1, 4, 5, 9 and 10).

We see that subject 6, 3, 9, 5 and 10 have consistent vocal discomfort patterns over the test days while subjects 2 and 7 have not. Subject 2 reported that he had a trainee on day 2 that did class instead of him while he had an intense voice use on day 1. Subject 7 indicated that there was a strike on day 2, she had less students than usual and reported less intense voice use on that day.

IV. Discussion

The adequacy of the device and the software was confirmed by high *wearability* and *easiness* scores, the subjects did not find the device bulky nor annoying to wear, the cursor was reported easy to move and to place on the right spot and it was easy to remember validating changed discomfort values.

The adequacy of the scale was confirmed by high scores at the questions evaluating the scale, it was regarded as highly adequate for the notation of vocal discomfort and the labels were rated as helpful.

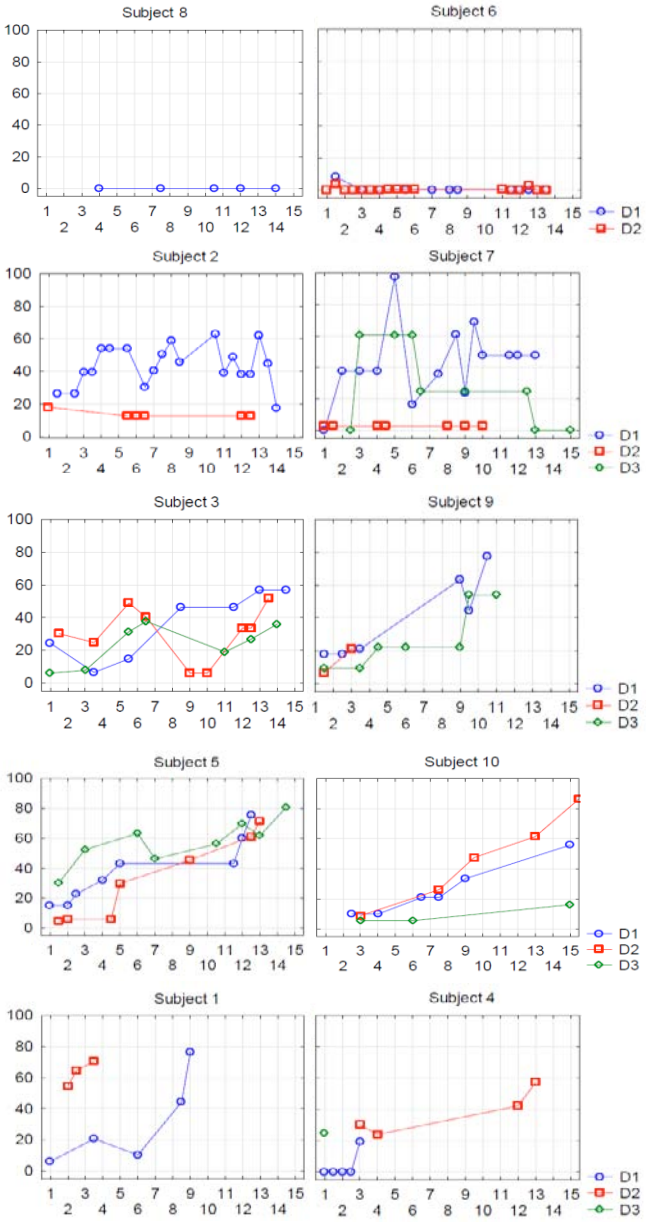


Fig. 6 Mean vocal discomfort values for each subject on each test day. Every half hour is plotted on the X-axis. Discomfort values are plotted on the Y-axis. (*D: Day.*)

The subjects reported that they seldom consciously omitted to report their vocal discomfort but they were more susceptible of forgetting to make the notations. Nevertheless, a mean of 10,4 notations were made with a mean interval of less than one hour. The adequacy of the scale was further supported by the overall increase in voice discomfort ratings over time, related to greater vocal load as the working day progressed. We stopped the recordings of the vocal discomfort at the end of the working days, in future studies where we will be able to follow subjects during a longer time, a decrease of the

vocal discomfort values would be expected with a reduced vocal load after work. This study was done in subjects who reported no dysphonia, it will be interesting in future studies to observe how these values are impacted by a vocal disorder.

Although the auditory prompt was not heard in a majority of cases, frequency of annotations was high and a reminder does not seem to be needed to obtain regular ratings over a day.

Seven days of data were lost due to subjects exiting the software, the implementation of a password controlled lock could prevent for that in the future.

Several subjects made spontaneous comments about the fact that the device had helped them to get conscious of their vocal use during a day and of their vocal discomfort, which they had not been reflecting over before their involvement in the study. This indicates that our software could be useful not only for diagnostic and outcome measures purposes but also in the context of vocal load monitoring in vocal professionals.

V. Conclusion

The real-time embedded tracking of patient reported vocal discomfort in professional settings can be advantageously performed by a portable device, embedding our auto-evaluation software. This study confirmed the validity of the scale we have developed for the tracking of changes in self-reported vocal discomfort in voice professionals.

References

[1] Lindstrom, F., Ohlsson, A.-C., Sjöholm, J., & Waye, K. P. Mean F0 Values Obtained Through Standard Phrase Pronunciation Compared With Values Obtained From the Normal Work Environment : A Study on Teacher and Child Voices Performed in a Preschool Environment. *J Voice* 2010:24:319-323

[2] Rantala, L., Vilkman, E., & Bloigu, R. Voice Changes During Work : Subjective Complaints and Objective Measurements for Female Primary and Secondary Schoolteachers. *J Voice* 2002:16:344-355.

[3] Södersten, M., Granqvist, S., Hammarberg, B., & Szabo, A. Vocal Behavior and Vocal Loading Factors for Preschool Teachers at Work Studied with Binaural DAT Recordings. *J Voice* 2002:16:356-371.

DETERMINANTS OF VOICE-RELATED SYMPTOMS AND COMPLAINTS IN DIFFERENT CATEGORIES OF TEACHERS : THE IMPORTANCE OF THE PSYCHO-EMOTIONAL COMPONENT

P.H. Dejonckere

The Institute of Phoniatrics, Utrecht University, NL, Federal Institute of Occupational Diseases, Brussels, B, Catholic University of Leuven, B.

Abstract: **Voice problems have become a major occupational health issue within the teaching community, as they frequently result in work absenteeism and need for professional re-orientation. Four main risk factors have been identified: voice loading, general health condition, environmental factors and psycho-emotional factors (occupational stress and frustration).**

In order to specifically consider the 'stress' aspect, we investigated voice complaints and voice-related quality of life in the teachers of a special education setting: the national military academy for future non-commissioned officers, actually adolescents in the age 12 to18. The outcomes were compared with those from recent reports about similar studies in common secondary schools in different European countries and in the USA. Our results demonstrate that the specific military teacher's population considered in this study clearly shows significantly lower prevalence of voice problems than comparable teacher's populations in 'common' secondary schools.

On the other hand, we investigated two specific groups of teachers supposed to have a heavier physical voice load than classical teachers: teachers of physical education and swimming teachers (in secondary schools). Concerning these two classes of teachers, the clear overall similitude with classical teachers provides a strong argument to consider that vocal load and environment is not the sole – or by far the most important – cause of voice complaints.

Keywords: **Teachers, voice load, psycho-emotional complaints, stress.**

I. INTRODUCTION

Voice problems have become a major occupational health issue within the teaching community, as they frequently result in work absenteeism and need for professional re-orientation. Four main risk factors have been identified [1-3]: voice loading (amount of teaching hours weekly), general health condition (upper airway infections, allergy, hearing loss, gastro-esophageal reflux…), environmental factors (noise, room acoustics, etc.), and psycho-emotional factors (occupational stress and frustration). Several recent publications suggest that the latter could play an important role. Actually the "stress" factor seems to consist of two components: the fear for aggressions and violence [4] and the lack of adequate coping strategies [5]. A growing number of misbehaving pupils and an increase in the size of the classes could account for deterioration in the last years [6].

In order to specifically consider the 'stress' aspect, we investigated voice complaints and voice-related quality of life in the teachers of a special education setting: the national military academy for future non-commissioned officers, actually adolescents in the age 12 to18. The most obvious difference between these specific surroundings and a normal secondary school is the discipline constraint, due to selection of the pupils, strict internal regulations and punitive sanctions (exclusion). The outcomes were compared with those from recent reports about similar studies in common secondary schools in different European countries and in the USA. Our working hypothesis is that enhanced discipline reduces stress in the teachers.

On the other hand, we investigated 2 specific groups of teachers supposed to have a heavier physical voice load than classical teachers: teachers of physical education and swimming teachers (in secondary schools).

The literature is however controversial: E.g. [7] found evidence for a higher risk to develop voice problems in teachers of physical education while [8] found no difference, although the teachers of physical education reported shouting much more (also in open air) compared to the other types of teachers.

In the current study, a comparison has been made among 3 matched groups: classical teachers, teachers of physical education and swimming teachers , in the same secondary schools.

II. METHODS

The basic tool for these studies was a questionnaire, including the Voice Handicap form, a validated, widely spread instrument for quantifying voice related quality of life, to be filled in by all teachers. Beside the VHI, the questions pertained to:

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

- general information (gender, age ….)
- health condition (general complaints / smoking & drinking habits / hearing / airway / reflux etc…)
- detailed information about career and teaching conditions
- a Yes/No statement: " In general, is your voice for you a problem ?"
- the Voice Handicap Index [9].

The questionnaire had to be filled in anonymously.

III. Results

A. Military teachers

73 fully completed questionnaires were suitable for analysis. 28 were incomplete. The response ratio was 70%. All were males.

(i) Prevalence of voice problems:
One single teacher (1%) gave a positive answer on the general Yes/No statement. Reference values for teachers are 52% [10], >55% [11] 59% [3], 43% [12], 54% [13] and for the general population 29% [14, 15] and 5% [11].

(ii) The median value for the VHI-score in our study was 5 (percentile 25: 2,8 and percentile 75: 10,3). The normative values (Median value, percentiles 25 & 75) are:
- Working population without occupational voice use: 5 (2-10)
- General population: 6 (2-12)
- Voice professionals: 7 (2-13)
- Teachers secundary school: 8 (3-15)

[16, 11, 3, 17, 18, 19, 20, 21].

B. Physical education teachers and swimming teachers, compared to classical teachers of the same schools

For this experiment 176 completed questionnaires were collected from teachers (physical education teachers and swimming), and 27 from swimming teachers. The response rate was 86%. The questionnaires of 59 healthy class teachers and 92 healthy physical education teachers were available for statistical analysis.

The median age of the classical teachers was 44,0, with a median of 16,0 working hours per week and 12,0 working years in education there were 63% male teachers and 37% female teachers.

The physical education teachers had a median age of 37,5; 21,0 working hours a week and the average of their years working as a physical education teacher was 13,7 . There were 59% male physical education teachers and 41% female teachers.

Analysis of results fails to show any difference between the three groups for the total VHI score, but a possible bias for swimming teachers is that they work in average a significant lower amount of hours per week. For the physical education teachers no significant difference of the VHI-total is found between the male (median 9,0) and the female teachers (median 10,0) (p = 0.17).

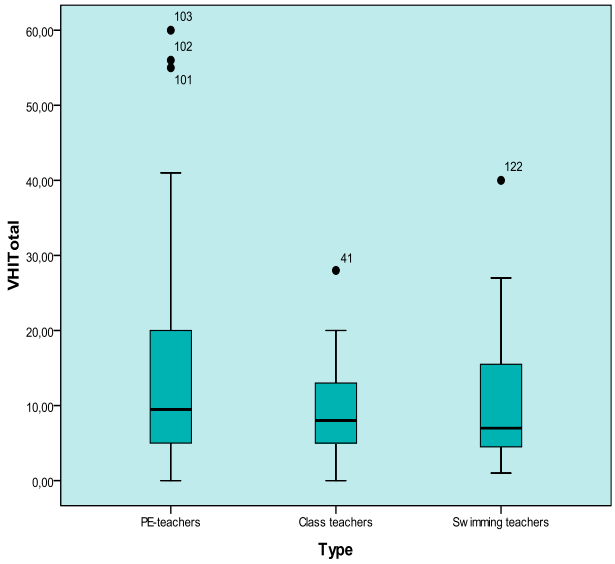


Figure 1: Boxplot of the VHI scores of the physical education teachers, the classical teachers and the swimming teachers

IV. Discussion and Conclusions

As far as comparative values are available, they indicate that our group of military teachers does not differ – as a general rule - in a biasing sense from the general population of secondary school teachers in the Netherlands or Belgium. The average age of our sample is 40 +/- 8,8 years, and the average age of all teachers in the Netherlands is 45. In our sample there are 77 % males and 23 % females, while in general in the Netherlands the proportions are 64 and 36 % for the secundary school teachers [22].

The voice loading for the military teachers and the environmental factors were – as far as possible - controlled, and appear to be not more favorable in the military academy than in a normal school. In our sample 95% of the teachers worked full-time while in Flanders this % is 65. In our sample , the average duration of the teaching career was 14,1 +/- 8 years. There were 25 – 30 pupils per class. All classrooms were visited and found quiet (the campus is at distance of town and highroad).

The general health condition is difficult to compare with the general teaching populations, as adequate statistics are lacking, but it could be that military teachers

– being for a part themselves military – have a better general condition than their colleagues from 'common' secundary schools. An indicator is that 58% of our military teachers actively practice sport. This is a possible bias. However, in our sample 40 % of the teachers were currently smokers.

Our results demonstrate that the specific teacher's population considered in this study clearly shows significantly lower prevalence of voice problems than comparable teacher's populations in 'common' secondary schools. Further, the psycho-social impact of voice problems considerably differs from what is known about secondary school teachers in general. The VHI scores of the military teachers are comparable to those of normal subjects without occupational voice use, and lower than those of the general population.

The specific surroundings and particularly the discipline context of the military academy seem to considerably reduce the stress related to teaching activities.

Except that the aspect "general health condition" should be investigated more in depth, as a possible partial bias, this study supports the hypothesis that psycho-emotional factors and occupational stress play an important role as risk factor for voice problems in teachers.

Concerning the physical education teachers and the swimming teachers, the clear overall similitude with classical teachers provides a strong argument to consider that vocal load and environment is not the sole – or by far the most important – cause of voice complaints.

REFERENCES

[1] Simberg S. Prevalence of vocal symptoms and voice disorders among teacher students and teachers and a model of early intervention. PhD Thesis. University of Helsinki, Hakapoino Oy. 2004.

[2] Kooijman P G C. The voice of a teacher: A multidimensional and dynamic process. Ph.D. Thesis. University of Nijmegen. 2006.

[3] Kooiijman PGC, Thomas G, Graamans K, de Jong FICRS. Psychosocial impact of the teachers voice troughout the career. J Voice 2007; 21: 316-324.

[4] Duffy OM, Hazlett DE. The impact of preventive voice care programs for training teachers: a longitudinal study. J. Voice 2004; 18: 63-70.

[5] De Jong FICRS, Cornelis BE, Wuyts FL, Kooijman PGC, Schutte HK, Oudes MJ, Graamans K. A psychosocial cascade model for persisting voice problems in teachers. Folia Phoniatrica et Logopaedica 2003; 55: 91-101.

[6] Simberg S, Sala E, Vehmans K, Laine A. Changes in the prevalence of vocal symptoms among teachers during a twelve year period. Journal of Voice 2005; 19: 95-102.

[7] Smith, E., J. Lemke, M. Taylor, H. Lester Kirchner and H. Hoffman. Frequency of voice problems among teachers and other occupations. *Journal of Voice*. 1998; 12; 480-488.

[8] Thibeault, S.L., R.M. Merrill, N. Roy, S.D. Gray and E.M. Smith. Occupational risk factors associated with voice disorders among teachers. *Annals of Epidemiology.* 2004 14; 786-792.

[9] Jacobson BH, Johnson A, Grywalsky C, Silbergleit A, Jacobson G, Benninger MS, Newman CW. The Voice Handicap Index (VHI): development and validation. American Journal of Speech and Language Pathology. 1997; 6: 66-70.

[10] Sapir S, Keidar A, Mathers-Schmidt B. Vocal attrition in teachers: survey findings. European Journal of Disordered Communication 1993; 28: 177-185.

[11] De Jong FICRS, Kooijman PGC, Thomas G, Huinck WJ, Graamans K, Schutte HK. Epidemiology of voice problems in Dutch teachers. Folia Phoniatrica et Logopaedica, 2006 ; 58: 186-98.

[12] Smith E, Gray SD, Dove H, Lester Kirchner H, Heras H (1997). Frequency and effects of teachers' voice problems. Journal of Voice,1997; 11, 81-87.

[13] Sala E, Laine A, Simberg S, Pentti J, Suonpaa J (2001). The prevalence of voice disorders among day care center teachers compared with nurses: a questionnaire and clinical study. Journal of Voice 2001; 15: 413-23.

[14] Roy N, Merrill RM, Thibeault S, Parsa RA, Gray SD, Smith EM. Prevalence of voice disorders in teachers and the general population. Journal of Speech, Language and Hearing Research 2004; 47, 281-293.

[15] Roy N, Thibeault S, Gray SD, Smith EM. Voice disorders in teachers and the general population: effects on work performance, attendance and future career choices. Journal of Speech, Language and Hearing Research, 2004; 47: 542-551.

[16] De Bodt M, Jacobson S, Musschoot S, Zaman S, Heylen L, Mertens F, Van de Heyning P, Wuyts F . De voice Handicap Index. Het kwantificeren van de psychosociale consequenties van stemstoornissen. Logopedie en foniatrie, 2001; 6: 159-162.

[17] Maertens K , de Jong FICRS. The Voice Handicap Index as a tool for assessment of the biopsychosocial impact of voice problems. B-ENT (Belgian Journal of Ear, Nose and Throat), 2007 ; 3: 61-66.

[18] Van Gogh CDL, Mahieu H, Kwik D, Rinkel R, Langendijk J, Verdonck-de Leew I .Voice in early glottic cancer compared to benign voice pathology. European Archives of Oto-Rhino-Laryngology, 2007; 264:1033-1038.

[19] Nawka T, Wiesman U, Gonnerman V .Validierung des Voice Handicap Index (VHI) in der deutschen Fassung. HNO, 2003; 51: 921-929.
[20] Guimares I, Abberton E. An investigation of the Voice Handicap Index with speakers of the Portuguese: preliminary data. J. Voice, 2004; 18: 71-82.
[21] Rosen CA, Lee AS, Osborne D, Zullo T, Murro T (2004). Development and validation of the Voice Handicap Index-10. Laryngoscope 2004 ; 114:1549-1556.
[22] Centraal Bureau voor Statistiek (Central Statistics Office). Jaarboek onderwijs in cijfers. Kluwer, Deventer (Nederland) 2007.

Special Session:
Acoustic analysis of Parkinsonian speech: issues methods and applications

Chairperson and introduction:

S. Sapir

SPECIAL SESSION

ACOUSTIC ANALYSIS OF PARKINSONIAN SPEECH: ISSUES METHODS AND APPLICATIONS

Shimon Sapir, Sabine Skodda, Athanasios Tsanas, Jan Rusz.

Parkinson's disease (PD) is a slowly progressive and highly debilitating disease of the central nervous system, affecting 8,000,000 or more people the world over. By the time the disease is diagnosed, 60% of nerve cells in the substantia nigra are degenerated and 80% of dopamine is depleted in the striatum. There is an urgent need for cost-effective methods to detect the disease in its early phases, to differentiate it from other diseases, and to monitor its progression and its response to treatment. Parkinsonian speech is characterized by abnormally low voice intensity, with vocal decay, poor voice quality, reduced prosodic pitch and loudness inflection, imprecise vowels and consonants, dysrhythmia and short rushes of speech, mumbling, and reduced speech intelligibility

Acoustic analysis of Parkinsonian speech is noninvasive, precise, valid, reliable and cost effective. Recently, there have been new acoustic analysis methods to capture different aspects of these speech abnormalities. The purpose of this seminar is to describe these methods, present empirical findings, and discuss the advantages, disadvantages, and potential solutions of these methods. Ultimately, a combination of these and other new methods is likely to yield a powerful way to detect early signs of PD, and to characterize and monitor the disease as it progresses or in response to treatment. These issues are addressed by the following presenters:

Shimon Sapir, Ph.D.

Sapir S, Ramig LO, Spielman JL, Fox C. (2010). **Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech.** J Speech Lang Hear Res. 53(1):114-25.

Sapir S, Spielman JL, Ramig LO, Story BH, Fox C. (2007). **Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings**. J Speech Lang Hear Res. 50(4):899-912.

Sapir S, Ramig L, Fox C. (2008). **Speech and swallowing disorders in Parkinson disease.** Curr Opin Otolaryngol Head Neck Surg. 16(3):205-10

Sabine Skodda, MD.

Skodda S, Flasskamp A, Schlegel U. (2011). **Instability of syllable repetition in Parkinson's disease-Influence of levodopa and deep brain stimulation.** Mov Disord. 26(4):728-30.

Skodda S, Flasskamp A, Schlegel U. (2011). **Instability of syllable repetition as a marker of disease progression in Parkinson's disease: a longitudinal study.** Mov Disord. 26(1):59-64.

Skodda S, Visser W, Schlegel U. (2010). **Vowel Articulation in Parkinson's disease.** J Voice. . [Epub ahead of print]

Skodda S, Rinsche H, Schlegel U. (2009). **Progression of dysprosody in Parkinson's disease over time--a longitudinal study.** Mov Disord. 24(5):716-22.

Athanasios Tsanas, doctoral student.

Tsanas A, Little MA, McSharry PE, Ramig LO. (2011).**Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity.** J R Soc Interface. 8(59):842-55.

Tsanas A, Little MA, McSharry PE, Ramig LO. (2010). **Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests.** IEEE Trans Biomed Eng. 57(4):884-93.

Jan Rusz, M.Sc., doctoral student.

Rusz J, Cmejla R, Ruzickova H, Klempir J, Majerova V, Picmausova J, Roth J, & Ruzicka E. (2011). **Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test.** Mov Disord. 2011 Apr 11. [Epub ahead of print]

Rusz J, Cmejla R, Ruzickova H, & Ruzicka E. (2011). **Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease**. J Acoust Soc Am, 129(1):350-367.

ROBUST PARSIMONIOUS SELECTION OF DYSPHONIA MEASURES FOR TELEMONITORING OF PARKINSON'S DISEASE SYMPTOM SEVERITY

Athanasios Tsanas[1,2], Max A. Little[2,3], Patrick E. McSharry[1,2,4], Lorraine O. Ramig[5,6]

[1] Systems Analysis, Modelling and Prediction group, Department of Engineering Science, University of Oxford, UK

[2] Oxford Centre for Industrial and Applied Mathematics (OCIAM), Mathematical Institute, University of Oxford, UK

[3] Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

[4] Smith School of Enterprise and the Environment, University of Oxford, UK.

[5] Speech, Language, and Hearing Science, University of Colorado, Boulder, Colorado, USA

[6] National Center for Voice and Speech, Denver, Colorado, USA

Abstract: **Parkinson's disease (PD) symptom severity is typically quantified using the standard clinical metric Unified Parkinson's Disease Rating Scale (UPDRS) which spans the range 0-176 (0 denotes healthy). This assessment requires the patient's physical presence in the clinic, is time consuming, and relies on the clinical rater's subjective evaluation and experience; practice has shown that expert clinicians might differ by as much as 4-5 UPDRS points in their evaluations. We had previously developed a statistical machine learning framework which enables accurate and objective quantification of average PD symptom severity using exclusively speech signals. For this purpose, we evaluated 132 speech signal processing algorithms (dysphonia measures), which attempt to capture distinctive characteristics in PD subjects' voice. On a very large database of about 6,000 phonations, we could replicate the clinical experts' assessments within less than two UPDRS points' error. In this paper, we focus on identifying the most successful of the original 132 dysphonia measures in estimating UPDRS using five robust feature selection techniques. We demonstrate that we can improve on our previous findings using only 15 dysphonia measures, where the selected measures also tentatively indicate the most representative pathophysiological characteristics in male and female PD voices.**

Keywords: **Parkinson's disease, telemedicine, Unified Parkinson's Disease Rating Scale, feature selection**

I. INTRODUCTION

Parkinson's disease (PD) is a crippling neurodegenerative disorder and it is estimated that more than one million people in North America alone are affected [1]. Reported incidence rates vary, but are in the range 10-20/100,000 [2]. Age is the single most important risk factor [2], and since the population is growing older, rates can be expected to rise further in the years to come. PD symptom monitoring requires the subject to make frequent physical visit to the clinic, and the dedicated time of expert clinicians in order to assess their general condition. Using a range of empirical physical tests, clinicians *subjectively* evaluate and map symptoms to a widely used metric known as the Unified Parkinson's Disease Rating Scale (UPDRS). This scale ranges from 0-176, where 0 denotes symptom-free and 176 total disability.

Building on recent evidence linking PD progression with vocal performance degradation [3], we have developed a statistical machine learning framework exploring the relationship between speech patterns and UPDRS [4-7]. In our studies we have used a collection of widely used speech signal processing algorithms (*dysphonia measures*) and have proposed a few novel algorithms [5], [8]. In this study, we review the most successful algorithms for estimating average PD symptom severity, and investigate plausible physiological explanations for this result. In order to decide on the most successful dysphonia measures out of the 132 from our most recent study [5], we use five new *feature selection* methods for choosing the best subset of dysphonia measures.

II. DATA

We use the speech-PD database originally presented by Goetz et al. [9], which we later summarized in [4] and [5]. This database was collected as part of a large clinical trial which involved a purpose-built *telemonitoring* device developed by Intel Corporation called *At-Home Testing Device* (AHTD). Various data, including tremor and dexterity tests, are involved, but we focus exclusively

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

on speech signals. We processed 5875 sustained vowel "ahh…" phonations from 42 people with PD who were diagnosed up to five years at trial onset, and who were followed up for a period of six months. All subjects remained non-medicated during the AHTD trial, and each subject was asked to complete the tests weekly. We have six phonations for each test session: four at a comfortable speaking level, and the remaining two at twice the initial loudness level (but without shouting). The phonations were recorded in the subjects' homes with a head mounted microphone. The voice signals were sampled at 24 kHz with 16 bits resolution and were recorded directly to a USB memory stick attached to the AHTD. Clinical UPDRS assessments were available at the trial baseline, after three months and six months. Supported by clinical findings which suggest that PD progression in early non-medicated subjects is approximately linear [10], to derive weekly UPDRS scores we used straightforward piecewise linear interpolation going exactly through the measured UPDRS values [4-7]. Further details regarding the device and the data can be found in [9] and [5].

Based on previous findings that males and females have distinct vocal PD characteristics [5], we investigate data from males (4010 phonations) and data from females (1865 phonations) separately.

III. METHODS

We aim to characterize the PD speech signals using the 132 dysphonia measures of [5], select the most useful dysphonia measures using a range of feature selection (FS) algorithms, and map the selected subset to UPDRS. Of particular interest is to identify the features that are common to the majority of the FS schemes, and attempt to decipher what pathophysiological PD characteristics these features reveal.

A. Computation of features

We compute 132 dysphonia measures for each phonation, using all the algorithms described in [5]. All algorithms are presented in detail in that study, and have been written using the Matlab software package. To summarize, the 132 features computed include 30 *jitter variants* (perturbations of the fundamental frequency), 21 *shimmer variants* (perturbations of the amplitude), 42 Mel Frequency Cepstral Coefficients (MFCCs), 8 fundamental frequency (F0) related measures (standard deviations, and differences of the fundamental frequency of each speech signal from the corresponding average fundamental frequency of age- and gender-matched healthy controls), 4 harmonics-to-noise (HNR) and noise-to-harmonic (NHR) ratios, and 27 additional dysphonia measures. These 27 dysphonia measures leverage exploit concepts related to: inability to sustain prolonged phonation, departure from periodicity in glottal opening and closure times, increased signal to noise ratio, and vocal differences compared to the average age- and gender-matched population.

Elsewhere, we have introduced the *Recurrence Period Density Entropy* (RPDE), *Detrended Fluctuation Analysis* (DFA) and *Pitch Period Entropy* (PPE) [8], and also *Vocal Fold Excitation Ratio* (VFER) family of measures, the *Empirical Mode Decomposition Excitation Ratio* (EMD-ER) family and the *Glottal Quotient* (GQ) [5]. RPDE quantifies the uncertainty in the measurement of the pitch period. DFA quantifies the stochastic self-similarity of turbulent noise in the speech signal. PPE quantifies the inefficiency of voice F0 control. The VFER family focuses on glottal pulses and forms signal-to-noise ratio measures, relying on energy, the Teager-Kaiser Energy Operator (TKEO), and entropy concepts. A similar concept motivates the EMD-ER family, which initially decomposes the speech signal into constituent, time-varying frequency-like components. The first components contain high frequencies in the signal (roughly attributable to noise), whereas the latter components represent power-like quantities. The GQ measures attempt to form a more sensitive *jitter-type* approach, relying on actual glottal cycle perturbation rather than F0 perturbation.

B. Feature selection

A common problem in applications with many features is the *curse of dimensionality*: a reduced feature subset may enhance the learner's performance and also promotes *model interpretability* [11]. That is, a reduced feature subset may enable insight into the underlying mechanisms of the system. Previously, we used two FS algorithms: the Least Absolute Shrinkage and Selection Operator (LASSO) and the elastic net [11]. Here, we complement these findings using five additional FS algorithms. Our rationale is to identify the features which are consistently selected and which with very high probability are the most useful in this application.

We used the following additional FS algorithms: 1) minimum Redundancy, Maximum Relevance (mRMR) [12], 2) the importance score of the Random Forest learner [13], 3) the ReliefF algorithm [14], 4) Information Gain (IG) [15] and 5) Sparse Bayesian Multinomial Logistic Regression (SBMLR) [16]. Investigation of the specific properties of each of the FS algorithms is beyond the scope of this study.

We run each FS algorithm in a 10-fold cross-validation setting: we randomly select 90% of the available data and based on this dataset we select features; the process is repeated a total of 10 times, each time with a new 90% randomly selected data points. Then, for each FS algorithm we select the subset which appeared most often in the 10 repetitions. This is the output of each FS scheme, and this output is used in the subsequent

Table 1: Selected dysphonia measures from five feature selection algorithms for males. The resulting out of sample mean absolute UPDRS estimation error (last row) uses these 15 features as input into the Random Forest learner.

LASSO	mRMR	RF	ReliefF	IG	SBMLR
6th MFCC	$VFER_{NSR,TKEO}$	DFA	6th MFCC	$VFER_{NSR,TKEO}$	8th MFCC
8th MFCC	6th MFCC	7th MFCC	DFA	$F_{0,Sun} - F_{0,exp}$	6th MFCC
$VFER_{SNR,TKEO}$	7th MFCC	6th MFCC	5th MFCC	0th MFCC	7th MFCC
$VFER_{mean}$	8th MFCC	4th MFCC	7th MFCC	DFA	9th MFCC
8th delta MFCC	10th delta MFCC	$VFER_{NSR,TKEO}$	8th MFCC	6th MFCC	3rd MFCC
12th delta MFCC	1st MFCC	2nd MFCC	3rd MFCC	$F_{0,mix} - F_{0,exp}$	4th MFCC
0th MFCC	3rd MFCC	8th MFCC	4th MFCC	$F_{0,Rapt} - F_{0,exp}$	5th MFCC
2nd MFCC	Log energy	3rd MFCC	9th MFCC	IMF_{energy}	10th MFCC
3rd MFCC	$VFER_{SNR,TKEO}$	Log energy	10th MFCC	$F_{0,Praat} - F_{0,exp}$	2nd MFCC
2nd delta MFCC	5th MFCC	1st MFCC	11th MFCC	8th MFCC	$IMF_{NSR,TKEO}$
3rd delta MFCC	$stdF0_{Praat}$	9th MFCC	Log energy	$VFER_{entropy}$	HNR_{std}
Std $F_{0,Sun}$	11th MFCC	11th MFCC	12th MFCC	3rd MFCC	11th MFCC
9th MFCC	HNR_{std}	$F_{0,Sun} - F_{0,exp}$	2nd MFCC	Log energy	12th MFCC
7th MFCC	4th MFCC	$F_{0,mix} - F_{0,exp}$	$F_{0,mix} - F_{0,exp}$	$VFER_{mean}$	0th MFCC
4th delta MFCC	9th MFCC	$F_{0,Rapt} - F_{0,exp}$	$F_{0,Rapt} - F_{0,exp}$	7th MFCC	1st MFCC
2.43±0.24	1.91±0.18	1.52±0.15	1.49±0.14	2.16±0.21	1.72±0.18

Table 2: Selected dysphonia measures from five feature selection algorithms for females. The resulting out of sample mean absolute UPDRS estimation error (last row) uses these 15 features as input into the Random Forest learner.

LASSO	mRMR	RF	ReliefF	IG	SBMLR
Log energy	Std $F_{0,Rapt}$	1st MFCC	4th MFCC	Std $F_{0,Rapt}$	0th MFCC
Std $F_{0,Rapt}$	Log energy	4th MFCC	Log energy	6th MFCC	4th MFCC
10th MFCC	$VFER_{SNR,TKEO}$	Std $F_{0,Rapt}$	0th MFCC	$F_{0,Praat} - F_{0,exp}$	2nd MFCC
PPE	10th MFCC	0th MFCC	2nd MFCC	1st MFCC	Log energy
12th MFCC	12th MFCC	Log energy	Std $F_{0,Rapt}$	0th MFCC	$IMF_{SNR,energy}$
$IMF_{SNR,TKEO}$	$JitterF0_{TKEO,std}$	2nd MFCC	DFA	HNR_{mean}	Std $F_{0,Rapt}$
8th MFCC	3rd MFCC	Std $F_{0,mix}$	1st MFCC	NHR_{mean}	$IMF_{SNR,entropy}$
11th MFCC	1st MFCC	HNR_{mean}	Std $F_{0,mix}$	Std $F_{0,mix}$	$IMF_{NSR,TKEO}$
$IMF_{NSR,SEO}$	Std $F_{0,mix}$	6th MFCC	5th MFCC	Log energy	Shimmer %
GNE_{mean}	0th MFCC	NHR_{mean}	7th MFCC	$Jitter_{TKEO,mean}$	$VFER_{entropy}$
3rd delta MFCC	12th delta MFCC	12th MFCC	6th MFCC	$Jitter_{TKEO,std}$	Std $F_{0,mix}$
HNR_{std}	11th MFCC	10th MFCC	10th MFCC	3rd MFCC	HNR_{mean}
5th MFCC	4th MFCC	5th MFCC	PPE	$Jitter_{TKEO,5\text{-}95\ prc}$	$OQ_{std\ closed\ cycle}$
2nd delta MFCC	HNR_{mean}	8th MFCC	3rd MFCC	5th MFCC	Std $F_{0,Praat}$
$GNE_{SNR,TKEO}$	GNE_{mean}	$VFER_{SNR,TKEO}$	11th MFCC	2nd MFCC	1st MFCC
2.73±0.23	2.28±0.21	2.22±0.24	2.14±0.25	2.34±0.27	2.61±0.25

mapping stage. To facilitate comparison, we focus only on the first 15 choices of each FS algorithm.

C. Feature mapping and generalization

As in our previous study [5], we use the Random Forests (RF) learner, mapping the dataset comprising the selected feature subset to the response (UPDRS). For details of RF we refer to Hastie et al. [11]. The generalization performance of the model is estimated using 10-fold cross validation, with 100 repetitions for statistical confidence. For each repetition we randomly permute the original data, and then use 90% of the data for training and the remaining 10% for testing. The error metric we minimize is the out of sample mean absolute error (MAE): $MAE = 1/N \sum_{i \in Q} |\hat{y}_i - y_i|$, where $\hat{y}_i$ is the predicted UPDRS and y_i is the actual UPDRS for the *i*th entry in the training or testing subset, *N* is the number of phonations in the training or testing subset, and Q contains the indices of that set. MAEs over the 100 cross-validation repetitions were averaged.

IV. RESULTS

We summarize the results of the FS algorithms in Table 1 for males, and Table 2 for females. Then, we use these selected feature subsets as input to the RF learner and compute the out of sample MAE, which is expressed in the form mean ± standard deviation (last row in Tables 1 and 2). Interestingly, the lowest MAE is given when the features provided by ReliefF or RF are input to the RF learner. As in [5], it appears that UPDRS in males can be estimated more accurately.

The choices of the FS schemes are interesting: overall, there is good agreement on the most useful features, suggesting we can be confident that these features may be the most representative of the PD pathophysiological status. It appears that MFCCs dominate in the male dataset, and F0-related measures dominate in the female dataset, verifying our previous findings [5]. Some of our recently proposed dysphonia measures are consistently selected across FS schemes; we elaborate on their properties in the following Section.

V. Discussion

We have used the dataset from [5] where 132 dysphonia measures were deployed in order to estimate UPDRS. We explored the data using five popular FS algorithms in order to determine the most useful feature subset for estimating UPDRS. In doing so, we have matched (female dataset) and also outperformed (male dataset) our previous results [5] using less than half the features. More importantly for our purposes, reducing the number of features promotes physiological understanding about what these dysphonia algorithms are measuring in PD. The present findings strongly reinforce our previous finding [5] that it may be beneficial to partition the data according to gender.

For males, the most important features appear to be the mid-range MFCCs, DFA, and $VFER_{NSR,TKEO}$. The MFCCs have traditionally appeared in speaker identification applications and have only relatively recently been introduced in the study of dysphonias [17]; our findings strongly support their use for monitoring Parkinson's disease symptom progression. This finding indicates that it is probably necessary to focus on formant resonances, as well as F0 and amplitude. That DFA is consistently selected verifies that increased turbulent noise is a feature of male PD voices. $VFER_{NSR, TKEO}$ being selected indicates that it is interesting to look at different frequency bands and determine signal to noise ratios in these bands; in particular, our experiments suggest that it may be useful to characterize frequencies above 2.5 kHz as 'noise', and frequencies below this as 'signal', in order to define *signal to noise* ratios [5].

For females, we note that in addition to MFCCs, F0-related measures are often selected. This perhaps stems from the physiological observation that normal vibrato is exacerbated in low fundamental frequency voices (that is, males) [18]. Although robust dysphonia measures to normal vibrato have been proposed (higher order jitter measures, PPE), we speculate that these approaches can only guard against low physiological tremor. This could suggest that vocal vibrato in females might be effectively removed using these robust vibrato-removal approaches, whereas the (comparatively larger) vibrato in males may not. Thus, robust F0 perturbation measures could indicate vocal pathology in females which might be otherwise overshadowed in males due to increased normal vibrato.

Acknowledgment

We are grateful to Mike Deisher, Bill DeLeeuw and Sangita Sharma at Intel Corporation. We also want to thank James McNames, Lucia M. Blasucci, Eric Dishman, Rodger Elble, Christopher G. Goetz, Andy S. Grove, Mark Hallett, Peter H. Kraus, Ken Kubota, John Nutt, Terence Sanger, Kapil D. Sethi, Ejaz A. Shamim, Helen Bronte-Stewart, Jennifer Spielman, Barr C. Taylor, David Wolff, and Allan D. Wu, who were responsible for the design and construction of the AHTD device and organizing the trials in which some of the data used in this study was collected.

References

[1] A.E. Lang, A.M. Lozano. "Parkinson's disease – First of two parts", *New England Journal Medicine*, 339, 1044-1053, 1998

[2] R. Pahwa, and E. Lyons. (Eds.) *Handbook of Parkinson's Disease*, 4th edition, Informa Healthcare, USA, 2007

[3] S. Skodda, H. Rinsche, U. Schlegel: "Progression of dysprosody in Parkinson's disease over time – A longitudinal study", *Movement Disorders*, Vol. 24 (5), pp. 716-722, 2009

[4] A. Tsanas, M.A. Little, P.E. McSharry, L.O Ramig, "Accurate telemonitoring of Parkinson's disease progression using non-invasive speech tests", *IEEE Transactions on Biomedical Engineering*, Vol. 57, pp. 884-893, 2010

[5] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity", *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011

[6] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression", IEEE Signal Processing Society, *International Conference on Acoustics, Speech and Signal Processing* (ICASSP '10), 594-597, 2010

[7] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig, "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity", *International Symposium on Nonlinear Theory and its Applications* (NOLTA), pp. 457-460, Krakow, September 2010

[8] M.A. Little, P.E. McSharry, E.J. Hunter, J. Spielman, L.O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", *IEEE Transactions on Biomedical Engineering*, Vol 56(4), pp. 1015-1022, 2009

[9] C.G. Goetz, G.T. Stebbins, D. Wolff, et al. "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device", *Movement Disorders*, Vol. 24 (4), 551-556, 2009

[10] M.W.M. Schüpbach, J-C. Corvol, V. Czernecki, M.B. Djebara, J-L. Golmard, Y. Agid and A. Hartmann, "The segmental progression of early untreated Parkinson disease: a novel approach to clinical rating", *J. Neurol. Neurosurg. Psychiatry*, Vol. 81, pp.20-25, 2009

[11] T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd ed., 2009

[12] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226-1238, 2005

[13] L. Breiman, Random Forests, *Machine Learning*, 45, 5-32 2001

[14] K. Kira and L.A. Rendell, "A practical approach to feature selection", *Proceedings of the Ninth International Conference on Machine Learning*, 249-256, 1992

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, Wiley, 1991.

[16] G.C. Cawley, N.L.C. Talbot and M. Girolami, "Sparse multinomial logistic regression via Bayesian L1 regularisation". In: *Advances in Neural Information Processing Systems*. MIT Press, pp. 209-216, 2006

[17] J.I. Godino-Llorente, P. Gomez-Vilda, M. Blanco-Velasco, "Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters", *IEEE Transactions on Biomedical Engineering*, Vol. 53, pp. 1943-1953, 2006

[18] *Baken, R.J. and R.F. Orlikoff. Clinical measurement of speech and voice*, San Diego: Singular Thomson Learning, 2nd ed., 2000

ACOUSTIC METRICS OF VOWEL ARTICULATION IN PARKINSON'S DISEASE: VOWEL SPACE AREA (VSA) Vs. VOWEL ARTICULATION INDEX (VAI)

S. Sapir[1], L. Ramig[2,3], J. Spielman[2,3], C. Fox[2,3]

[1]Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel

[2]Department of Speech, Language and Hearing Sciences University of Colorado, Boulder

[3]National Center for Voice and Speech-Denver

Abstract: **Acoustic analysis of speech is a powerful, noninvasive, and cost effective tool to study different aspects of motor speech disorders such as the dysarthria associated with PD. In this presentation we will discuss the rationale for using acoustic analysis, its advantages and disadvantages, and methods to overcome these disadvantages. As an example, we will address the use of vowel space area (VSA) in the study of dysarthric vowel articulation in PD. Although the VSA is theoretically driven, it is highly sensitive to inter-speaker variability, which, statistically speaking, introduces noise. This noise can mask important differences that do exist between speakers with and without PD. Some of this noise can be reduced by logarithmic transformation of the formant frequencies. However, even with this transformation, some statistical noise might be still present. Recently Sapir and colleagues introduced two acoustic metrics -- the Vowel Articulation Index (VAI) and its inverse, the Formant Centralization Ratio (FCR) -- that are theoretically driven and empirically tested. These metrics show promise as they effectively reduce inter-speaker variability noise while maintaining high sensitivity to vowel centralization (the latter reflecting abnormally reduced (hypokinetic) articulatory movements in PD). Data will be presented of 38 individuals with Parkinson's disease and 14 healthy controls whose speech was effectively differentiated by the VAI, but not the VSA, yet the logarithmically scaled VSA (LnVSA) did significantly differentiate between dysarthric and normal speech, although not as strongly as the VAI.**

Keywords: **Parkinson disease, acoustic analysis, speech**

I. INTRODUCTION

Individuals with Parkinson's disease (PD) often suffer from hypokinetic dysarthria, a neuromuscular disorder of voice and speech, resulting and characterized by reduced vocal loudness, monotone voice, and imprecise consonants and vowels.

Most types of dysarthria, including that associated with PD, are characterized by articulatory undershoot, i.e., reduced range of articulatory movements, to the extent that the intended place and degree of vocal tract constriction are not fully achieved. This undershoot is likely to result in vowel formant centralization; i.e., formants that normally have high center frequencies tend to have lower frequencies, and formants that normally have low center frequencies tend to have higher frequencies [1,2].

A common way to represent this centralization is with the VSA [3]. In English, the VSA is usually constructed by the Euclidean distances between the F1 and F2 (frequency) coordinates of the corner vowels /i/, /u/, and /a/ (triangular VSA), or the corner vowels /i/, /u/, /a/, and /ae/ (quadratic VSA) in the F1-F2 plane. The formula of the VSA constructed with the vowels /i/, /u/ and /a/, is ABS((F1i*(F2a–F2u)+F1a*(F2u–F2i)+F1u*(F2i–F2a))/2).

Due to articulatory undershoot and subsequent centralization of vowels, the VSA in the speech of individuals with dysarthria is expected to be compressed relative to that of normal speech (Kent & Kim, 2003). Improvement in speech due to natural recovery or treatment effects should be reflected in the expansion of the VSA toward normalcy (e.g., Sapir et al, 2003).

Although several studies demonstrated the ability of the VSA to differentiate between dysarthric and normal speech and to monitor treatment effects (e.g., Liu et al., Sapir et al, 2003), other studies failed to do so, even though a trend toward centralization of vowels was evident (e.g., Weismer et al., 2001). The reasons for the inconsistent performance of the VSA are not clear, although interspeaker variability appears to be a major factor. Interspeaker variability in vowel formant frequencies and VSA is expected due to numerous factors (cf. Sapir et al., 2010), the most obvious of which are anatomical differences, such as those associated with gender and age (re: size and shape of the vocal tract).

It is clear that to improve differentiation between dysarthric and normal speech, the acoustic metric must be minimally sensitive to interspeaker variability and maximally sensitive to vowel formant centralization.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Recently, Sapir introduced two acoustic metrics that are designed to be minimally sensitive to interspeaker variability and maximally sensitive to vowel formant centralization. These metrics include the Vowel Articulation Index (VAI), expressed as (F2i+F1a)/(F2u+F2a+F1u+F1i), and its inverse, the Formant Centralization Ratio (FCR), expressed as (F2u+F2a+F1u+F1i)/ (F2i+F1a). (Sapir et al., 2006; Sapir et al., 2010). Note that in the VAI the numerator is likely to decrease and the denominator is likely to increase with vowel formant centralization, whereas in the FCR the numerator is likely to increase and the denominator to decrease with vowel centralization. Importantly, at least in American English, the normal VAI values should be close to 1.0, as the sum of formant frequencies in the denominator is very similar to the sum of formant frequencies in the numerator. Thus, the VAI may be considered a function that normalizes the relationships between the vowels across speakers. The purpose of the present study is to demonstrate the ability of the VAI, VSA, and LnVSA, to differentiate between normal and abnormal vowel articulation. We predicted that the VAI will perform best and the VSA worst. We also predicted that the LnVSA will perform better than the VSA because logarithmic scaling of formant frequencies tend to reduce interspeaker variability.

II. METHODS

Subjects. The subjects in this study participated in our previous study (Sapir et al., 2010). They all spoke American English as their first language and the majority of them resided in Tucson Arizona or Denver Colorado. Of these individuals, 38 had idiopathic Parkinson's disease (PD) (19 M, 19 F) with dysarthria of different levels of severity, and 14 individuals (7 M, 7 F) served as healthy, age-matched and gender-matched controls (HC). The VAI and VSA were constructed from the frequencies of the first (F1) and second (F2) formants of the vowels /i/, /u/, and /a/. These frequencies were also logarithmically scaled for the construction of a logarithmic version of the VSA (henceforth, LnVSA). The vowels /i/, /u/, and /a/ were extracted from the phrases “The blue spot is on the key,” “The potato stew is in the pot” and “Buy Bobby a puppy” (target words: “key”, “stew”, and ”Bobby") or the phrase “The stew pot is packed with peas" (target words "stew", "pot", "peas"), with several repetitions of each of the phrases. Details of the recordings and acoustic analysis are described elsewhere (Sapir et al., 2010).

III. RESULTS

The main findings of this study are summarized in Table 1. The table shows the means and SDs for the two groups (PD, HC) and for three acoustic metrics (VSA, LnVSA, VAI), as well as t-test results and p values for significance. Also, the coefficient of variation (CV) for the two groups and three metrics (VSA, LnVSA, VAI) are shown at the bottom of the table. Effect size (ES) measures (Cohen, 1988) are also used to indicate the clinical significance of the differences between the two groups the degree. In general, a value of 0.80 and higher indicates highly significant differences between the two groups. A value of 0.50 a medium effect and a value of 0.20 indicates a small or a negligible effect. As can be seen, the VSA does not significantly differentiate between the two groups (PD vs. HC). The LnVSA improves performance considerably, whereas the VAI performs best, both statistically and in terms of a large effect size.

Table 1. The ability to differentiate between the dysarthric and normal vowel articulation by the VSA, LnVSA, and VAI. CV = Coefficient of Variation ; ES= Effect Size (>0.8 large, 0.5 medium, 0.2 small effect).

		VSA (Hz)	LnVSA (LnHz)	VAI
PD (n=38)	Ave=	232120	0.23	0.96
	SD=	(96155)	(0.08)	(0.08)
HC (n=14)	Ave=	279524	0.28	1.05
	SD=	(68810)	(0.07)	(0.08)
t-test	p=	0.0579	0.0099	0.0006
	ES=	0.57	0.77	1.24
PD	CV=	41%	33%	8%
HC	CV=	25%	24%	7%

IV. DISCUSSION

These findings strongly suggest that by reducing interspeaker variability and by maximizing sensitivity of the acoustic metric to the differences between normal and abnormal speech one can improve the reliability and validity of the acoustic analysis. Our task is to do the same for other acoustic metrics of speech.

Once we have acoustic metrics that comply with these two criteria, we can combine the different acoustic

metrics and use more sophisticated analyses to differentiate between normal and abnormal speech and to monitor changes associated with disease progression and treatment effects. Finally, we addressed only one issue related to improving speech signal processing for clinical and research processes. There are other important factors that we should consider, such as the problem of recording speech in a noisy environment, using inappropriate recording equipment and procedures, and choosing the wrong speech tasks to elucidate and register the speech abnormalities in PD.

V. CONCLUSIONS

Unlike the VSA, the VAI is a powerful acoustic metric to reduce interspeaker variability and enhance sensitivity to dysarthric vowel articulation.

REFERENCES

[1] Sapir S, Spielman JL, Ramig LO, Story BH, Fox C. (2007). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings. *J Speech Lang Hear Res*. 50, 899-912.

[2] Sapir S, Ramig LO, Spielman JL, Fox C. (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. *J Speech Lang Hear Res.* 53, 114-25.

[3] Kent, R., & Kim, Y. (2003). Toward an acoustic typology of motor speech disorders. *Clin Ling Phonetics*, 17, 427-45.

[4] Sapir, S., Spielman, J., Ramig, L., Hinds, S., Countryman, S., Fox, C., & Story, B. (2003). Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on ataxic dysarthria: a case study. *Am J Speech Lang Path,* 12, 387-99.

[5] Liu, H., Tsao, F., & Kuhl, P. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *J Acoust Soc Am* , 117, 3879-89.

[6] Weismer, G., Jeng, J-Y., Laures, J., Kent, R., & Kent, J. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniat Logopaed* , 53, 1-18.

[7] Sapir, S. (2006, Feb). *Effects of LSVT on speech articulation in dysarthric individuals with Parkinson's disease: Acoustic and perceptual correlates.* A paper presented at the Congress of the European Federation of Neurological Societies, Istanbul, Turkey

[8] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd Edition). Hillsdale, NJ: Earlbaum,

ACOUSTIC ANALYSIS OF SPEECH AS A PROMISING INSTRUMENT FOR MONITORING AND DIFFERENTIAL DIAGNOSIS OF PARKINSONS'S DISEASE

S. Skodda

Department of Neurology, Knappschaftskrankenhaus, Ruhr-University of Bochum, Germany

Abstract: **Parkinsonian speech is characterized by abnormally low voice intensity, with vocal decay, poor voice quality, reduced prosodic pitch and loudness inflection, imprecise vowels and consonants, dysrhythmia and short rushes of speech, mumbling, and reduced speech intelligibility. Recently, there have been new acoustic analysis methods to capture different aspects of these speech abnormalities. In this review, selected studies are summarized in order to illustrate the application of acoustic analysis of speech for the objective measurement and quantification of different aspects of Parkinsonian dysarthria.**

Keywords : **Parkinson´s disease, acoustic analysis of speech, hypokinetic dysarthria, dysprosody, vowel articulation, syllable repetition, motor speech performance**

I. INTRODUCTION

Parkinson´s disease (PD) is a neurodegenerative disorder characterized by progressive loss of dopaminergic neurons, primarily in the substantia nigra pars compacta, [1]. According to the Braak staging, PD begins as a synucleopathy in non-dopaminergic structures of the lower brainstem or in the olfactory bulb with subsequent rostral progression and affection of the substantia nigra [2]. The progressive dopaminergic loss is associated with a variety of motor and non-motor deficits in PD patients. In addition to the most ostensible symptoms as muscular rigidity, tremor, bradykinesia and postural instability, many patients develop a distinctive alteration of speech characterized as hypokinetic dysarthria. In a survey, the prevalence of dysarthria in PD was about 70% [3]. Dysarthria can emerge at any stage of the disease and worsen in the later stages [4, 5], causing a progressive loss of communication and leading to social isolation. Based upon the perceptual analysis of a large sample of dysarthric speakers, Darley, Aronson and Brown primarily defined a salient cluster of deviant speech dimensions in Parkinsonian dysarthria including a harsh breathy voice quality, reduced variability of pitch and loudness, reduced stress, imprecise consonant articulation and short rushes of speech interrupted by inappropriate periods of silence [6, 7]. Together, these features give hypokinetic dysarthria its distinctive gestalt of a flat, attenuated and sometimes accelerated quality [6, 7]. Logeman and colleagues established a general profile of hypokinetic dysarthria in a group of 200 PD patients, where almost 90% had voice disorders characterized by hoarseness, roughness, tremulousness and breathiness [8]. About half of the speakers featured articulatory problems, and 20% had speech rate abnormalities characterized by syllable repetitions, irregularities of syllable length and excessive speech pauses. According to this study, the authors supposed voice abnormalities to be the prominent attribute of hypokinetic dysarthria with the assumption of further subgroups including articulatory and speech rate deviations. In a further investigation on a large group of PD patients performed by Ho and colleagues, voice impairment was present even in the early stages of the disease with additional articulatory deficits and disturbance of fluency in the more advanced stages of PD [4]. Though, changes of speech rate and regularity were also observed in a subgroup of only mildly affected patients leading to the hypothesis that fluency deficits might be an isolated feature of hypokinetic dysarthria independent from voice and articulatory impairment [4].
Since this first systematic characterization of hypokinetic dysarthria, there has been a wealth of subsequent investigations based upon perceptual, acoustical and electrophysiological methods which further refined the description of speech disturbance in PD. However, although there is some evidence for a manifestation of hypokinesia and muscular rigidity of the vocal tract, there is still ambiguity concerning the pathophysiological mechanism of the different aspects of Parkinsonian speech disturbance in detail.
In this presentation it will be shown, that acoustic analysis of voice and speech in PD and related movement disorders may be a helpful instrument to gain further insight into the underlying pathophysiology by objective and quantifiable measurement of distinctive speech parameters and might therefore serve as a "window into the disease".

II. METHODS

In the different investigations, acoustic analysis of speech was based upon a standard reading task consisting of four complex sentences [9, 10, 11] or a simple syllable

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

repetition paradigm where participants were asked to repeat a single syllable in a self-chosen steady pace [12, 13]. Speech samples were digitally recorded and analyzed using the software PRAAT [14]. For the description of *intonation*, fundamental frequency variability (F_0SD) and fundamental frequency range were applied. Aspects of *speech velocity* and *fluency* were illustrated by total and net speech rate (TSR and NSR), pause ratio (PR%) and the fraction of intra-word pauses related to overall speech pauses (Pinw%). Furthermore, the acceleration of speech rate in the course of reading was defined as articulatory acceleration (AA). The vowel articulation index (VAI) first established by Roy and Sapir [15] was used for the measurement of *vowel articulation*. Concerning *syllable repetition capacity*, the relative coefficient of variance of syllable length (COV) was introduced as a measure of steadiness in the course of the performance.

Participants consisted of different samples of patients with PD, age- and gender-matched healthy speakers and – in one investigation – of patients with progressive supranuclear palsy (PSP). At the time of examination, all patients were under stable dopaminergic medication. Global motor impairment of all patients was rated according to the Unified Parkinson´s Disease Rating Scale (UPDRS) and Hoehn&Yahr stages.

III. Results

A. Progression of dysprosody in PD over time [9]

In a group of 50 patients with PD which were tested and re-tested after at least 12 months (mean 25 months) according to the reading task, TSR and NSR declined from first to second examination, especially in the male patients, but showed no significant differences to the control group. The course of pitch variation revealed some gender particularities. Whereas female patients' pitch variability declined over time, male patients' intonation variability remained relatively stable. F_0SD in male and female patients with PD were significantly reduced compared with the control group in the first examination and the follow up as well. Progression of prosodic impairment over time showed no correlation to disease duration or UPDRS motor score.

B. Vowel articulation in PD [10]

In a group of 68 patients with PD with mild dysarthria (1 point according to the "speech" item 18 of UPDRS) and 32 age-matched control persons, vowel articulation and speech rate were measured. F1 and F2 frequency values of the German vowels /a/, /i/, and /u/ were extracted from defined words within the reading text. Description of vowel articulation was based on measures of VAI. As main results, VAI values were significantly reduced in male and female PD patients as compared with the accordant control group. NSR was negatively correlated to VAI only in female PD speakers. No correlations were seen between vowel articulation and UPDRS and stage of disease. Obviously, some aspects of altered speech performance in PD seemed to feature some gender-specific patterns.

C. Acoustic analysis in PSP [11]

Based upon the reading task, 26 patients with PSP were examined in comparison to a group of 30 patients with PD. In the PSP group, NSR, F_0SD and Pinw% (as a measure of precision of consonant articulation) were significantly reduced, whereas %PR was prolonged as compared with the PD group. Only in the male PSP patients, vowel articulation was found to be impaired. Global speech performance – as rated by perceptual impression – was worse in the PSP group in comparison with the PD group and showed a correlation to some distinct speech dimensions obtained by acoustic analysis.

D. Stability of syllable repetition in PD [12]

Based upon the syllable repetition task, 73 patients with PD and 43 healthy speakers were tested concerning the capacity to steadily repeat a single syllable (/pa/) in a self-chosen isochronous pace. COV of interval length and the change in interval length with successive utterances were measured for the description of pace stability throughout the performance. Then, participants had to identify irregularities of 30 played-back audio tests. Patients with PD showed significant difficulties in steadily executing a syllable repetition task with a significant elevation of COV and showed a clear tendency to pace acceleration in the course of the performance. However, there were no differences in the correct auditory identification of rhythm irregularities between the PD group and controls. As compared to healthy controls, the PD group featured disabilities in performing a steady sequence of utterances, which cannot be explained solely by impaired acoustical feedback mechanisms. The pattern of pace disturbance showed similarities with the finding of speech acceleration and rhythm irregularity in the course of reading or more complex conversational speech and therefore might share the same pathophysiology.

E. Instability of syllable repetition in the course of the disease [13]

As previously shown, Parkinsonian speakers show a tendency to articulatory acceleration and have difficulties to keep the steady pace of repeated syllables. The aim of the subsequent study was to analyse the stability of motor speech performance based upon the syllable repetition

paradigm during the course of disease to find a potential marker of disease progression in PD. 58 patients with PD and 35 controls were tested and re-tested after at least 12 months (mean 33.40 months). In the PD group, motor impairment was similar at first and second visit. Participants had to repeat the syllable /pa/ in a self chosen steady pace. Besides the calculation of COV as a measure of instability of repetition, the "percental pace acceleration in the course of the performance" (%PA) was further introduced. Patients with PD showed a significant elevation of COV and %PA indicating an instability of syllable repetition and a tendency to pace acceleration in the course of performing. Furthermore, in the PD group, COV and %PA showed a significant deterioration from first to second examination. Instability of steady syllable repetition in PD showed characteristic changes during the course of the disease, but no correlation with general motor impairment.

IV. Discussion

The aforementioned investigations can serve as an example for the application of acoustic analysis of speech in PD. Since certain parameters of dysprosody and stability of syllable repetition feature distinct patterns of deterioration in the course of disease and seem to be independent from global motor impairment, these speech variables might have the potential to serve as marker of disease progression. Furthermore, vowel articulation as measured by VAI seemed to be impaired even in Parkinsonian patients with only mild dysarthria (when perceptually rated) and might therefore turn out to become a useful tool for the early detection of subclinical speech impairment in PD.

Instability of syllable repetition in PD might be interpreted as dysfunction of planning, preparing and executing basic motor speech sequences which share some similarities with the impaired execution of repetitive limb movements and therefore might indicate a shared pathophysiology.

In a small series of patients, acoustic analysis of several distinct speech variables was able to differentiate Parkinsonian speakers from patients with PSP. Since in PSP, the neuropathological changes are more widespread than in PD, comprising basal ganglia as well as pontine and further brainstem and sometime cerebellar regions, the resulting dysarthria in PSP is more severe and may include hypokinetic, spastic and ataxic components which might be detected by acoustic analysis of speech.

One main limitation of the presented investigations might be he the fact that all patients were under dopaminergic medication at the time of the examination and therefore, therapeutic or detrimental effects of the medication on the different speech variables cannot be ruled out. However, according to previous studies of our group, instability of syllable repetition on the one hand and several further speech parameters had shown no significant changes under short- and long-term dopaminergic stimulation [16, 17]. These findings justify the hypothesis that certain aspects of Parkinsonian dysarthria are independent from dopaminergic transmission.

V. Conclusion

According to the exemplified studies, acoustic analysis of speech in PD and related disorders might serve as a non-intrusive and easy applicable instrument for the measurement and monitoring of different speech dimensions. Furthermore, it might be helpful to generate and verify hypothesis about pathophysiological relations between speech and general motor performance in PD and might therefore serve as a "window into the disease".

References

[1] C.D. Marsden "Parkinson´s disease." *J. Neurol. Neurosur. Psychiatry,* vol. 57, pp. 672-81, 1994

[2] H. Braak , J.R. Bohl, C.M. Müller, O. Rüb , R.A. de Vos et al. "The staging procedure for the inclusion body pathology associated with sporadic Parkinson´s disease reconsidered." *Mov. Disord.,*; vol. 21, pp. 2042-51, 2006

[3] L. Hartelius, P. Svensson. "Speech and swallowing symptoms associated with Parkinson´s disease and multiple sclerosis: a survey" *Folia Phon. Logop.,* vol. 46, pp.9-17, 1994

[4] A. Ho, R. Iansek, C. Marigliani , J.L. Bradshaw, S. Gates. "Speech impairment in a large sample of people with Parkinson's disease." *Behav. Neurol*., vol. 11, pp. 131-37, 1998

[5] W.J. Mutch, A. Strudwick S.K. Roy, A.W. Downie. "Parkinson´s disease: disability, review, and management." *BMJ*, vol. 293, pp. 675-77, 1986

[6] F.L. Darley, A.E. Aronson, J.R. Brown. "Differential diagnostic patterns of dysarthria." *J. Speech Hear. Res.*, vol. 12, pp. 249-62, 1969

[7] F.L. Darley, A.E. Aronson, J.R. Brown. "Clusters of deviant speech dimensions in the dysarthrias." *J. Speech Hear. Res.,* vol. 12, pp. 462-96, 1969

[8] J.A. Logemann, H.B. Fisher, B. Boshes, E.R. Blonsky. "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinsonian patients." J. Speech Hear. Dis., vol. 43, pp. 47-57, 1978

[9] S. Skodda, H. Rinsche, U. Schlegel. „Progression of dysprosody in Parkinson´s disease over time – a

longitudinal study.“ *Mov. Disord.*, vol. 24, pp. 716-22, 2009
[10] S. Skodda, W. Visser, U. Schlegel. ”Vowel articulation in Parkinson´s disease.” *J. Voice*, 2010 (epub ahead of print)
[11] S. Skodda, W. Visser, U. Schlegel. ”Acoustic analysis of speech in progressive supranuclear palsy.” *J. Voice*, 2010 (epub ahead of print)
[12] S. Skodda, A. Flasskamp, U. Schlegel. ”Instability of syllable repetition as a model of impaired motor processing: is Parkinson´s disease a “rhythm” disorder ?” *J. Neural. Transm.,* vol. 117, pp. 605-12, 2010
[13] S. Skodda, A. Flasskamp, U. Schlegel. ”Instability of syllable repetition as a marker of disease progression in Parkinson´s disease: a longitudinal study.” *Mov. Disord.,* vol. 26, pp. 59-64, 2011
[14] P. Boersma, D. Weenik. “PRAAT: a system for doing phonetics by computer.” *Report of the Institute of Phonetic Sciences of the University of Amsterdam*, 1996, available at: http:// www.fon.humuva.nl/praat
[15] N. Roy, S.L. Nissen, C. Dromes, S. Sapir. ”Articulatory changes in muscle tension dysphonia : evidence of vowel space expansion following manual circumlaryngeal therapy.” *J. Commun. Disord.,* vol. 42, pp. 124-35, 2009
[16] S. Skodda, W. Visser, U. Schlegel. ”Short- and long-term dopaminergic effects on dysarthria in Parkinson´s disease.” J. *Neural. Transm*., vol. 117, pp. 197-205, 2010
[17] S. Skodda, A. Flasskamp, U. Schlegel. ”Instability of syllable repetition in Parkinson´s disease – influence of levodopa and deep brain stimulation.” *Mov. Disord.,* vol. 26, pp. 728-30, 2011

ACOUSTIC ANALYSIS OF VOICE AND SPEECH CHARACTERISTICS IN EARLY UNTREATED PARKINSON'S DISEASE

J. Rusz[1,2], R. Cmejla[1], H. Ruzickova[2], J. Klempir[2], V. Majerova[2], J. Picmausova[2], J. Roth[2], E. Ruzicka[2]

[1] Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Circuit Theory, Prague, Czech Republic

[2] Charles University in Prague, First Faculty of Medicine, Department of Neurology and Centre of Clinical Neuroscience, Prague, Czech Republic

Abstract: **Parkinson's disease (PD) is a neurological illness characterized by progressive lost of dopaminergic neurons, primarily in the substantia nigra pars compacta. Changes in speech associated with hypokinetic dysarthria are a common manifestation in patients with idiopathic PD. The aim of this study is to investigate the feasibility of automated acoustic measures for the identification of voice and speech disorders in PD. The speech data were collected from 46 Czech native speakers, 24 with early PD before receiving pharmacotherapy treatment. We have applied several traditional and non-standard measurements in combination with statistical decision-making strategy to assess the extent of vocal impairment of recruited speakers. Subsequently, we have applied support vector machine to find the best combination of measurements to differentiate PD from healthy subjects. This method leads to overall classification performance of 85%. Admittedly, we have found relationships between measures of phonation and articulation and bradykinesia and rigidity in PD. In conclusion, the acoustic analysis can ease the clinical assessment of voice and speech disorders, and serve as measures of clinical progression as well as in the monitoring of treatment effects.**

Keywords: **Parkinson's disease, speech disorders, hypokinetic dysarthria, acoustic analysis, biomedical application**

I. INTRODUCTION

Following the recent findings on the pathogenesis of Parkinson's disease (PD), increased interest has been paid to the nonmotor symptoms indicating an early affection of the lower brainstem that may precede the accession of the main motor signs of PD [1].

As a part of the nonmotor symptoms, voice and speech disorders are still considered to occur inconstantly and to tend to be nonspecific, making them of little diagnostic usefulness in early disease. On the other hand, previous research has shown that deficiencies in speech affect approximately 75-90% people with PD [2, 3]. The most salient features of PD speech impairment include deficits in the production of vocal sounds and motor involvement of articulation [3-5]. Moreover, it has been demonstrated that PD-related dysarthria can affect all different speech subsystems including *respiration*, *phonation*, *articulation*, and *prosody* [6, 7]. Patients with PD can manifest abnormalities related to all dimensions of speech including monoloudness, monopitch, imprecise articulation, variable speech rate, hoarseness, reduced stress, speech disfluencies, inappropriate silence, and others [7].

To clinically test voice and speech disorders, there are various vocal tests that have been proposed to assess the extent of these symptoms including sustained phonation, diadochokinetic (DDK) task (diadochokinesis connected with articulation), and variable reading of sentences or spontaneous speech [8, 9], that can be subsequently assessed with various traditional and novel acoustic measurements [10]. In our studies, we focus to characterize the speech and voice disorders in the early stages of PD, where the progression of speech symptoms is not affected by medication. In order to find PD-related speech features and separate patients with PD from healthy control (HC) persons, we use several traditional and novel acoustic measurement techniques as well as statistical learning or decision theory.

II. DATA

We used the database of PD speech recordings which we reported in [11]. From 2007 to 2009, a total of 46 Czech native participants were recruited for this research. 24 of these subjects (20 men and 4 women) fulfilling the diagnostic criteria for PD were examined immediately after the diagnosis was made and before the symptomatic treatment was started. As a control group, 22 persons (15 men and 7 women) with no history of neurological or communication disorders were included. None of the participants had been under voice therapy and all gave their consent to the vocal tasks and recording procedure.

The speech data were recorded in a quiet room with a low ambient noise level using an external condenser

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

microphone placed at approximately 15 cm from the mouth and coupled to a Panasonic NV-GS 180 video camera. The voice signals were sampled at 48 kHz, with 16-bit resolution; the video material was not used. All subjects were recorded at the time of a single session with a speech pathologist. Each participant was instructed to perform at least two times three vocal tasks including sustained phonation, diadochokinetic task, and running speech as a part of a larger protocol. Detailed description of the recording and data can be found in [11].

III. METHODS

We aimed to characterize the PD speech performance, applying the statistical decision-making theory to several acoustic measurements to explore how PD-related vocal symptoms differ from the speech performances of the wider norm of healthy speakers. Subsequently, we designed quick vocal test in order to reduce the time required for voice investigation, represent all speech subsystems, and create reliable assessment in practice, and tested performance of this test in separating PD subjects from HC participants. Finally, we search for a possible correlation of the voice parameters with respect to the duration and severity of disease.

A. Characteristics of voice and speech disorders

We have mainly focused on four speech subsystems including phonation, respiration, articulation, and prosody. For each of these subsystems, we computed several acoustic measures; all the algorithms are described in [10]. In examining phonation of PD speakers, we computed 4 features including *jitter* (the extent of variation of voice range), *shimmer* (the extent of variation of expiratory flow), *noise-to-harmonics* (NHR), and *harmonics-to-noise* (HNR) ratios (the amplitude of noise relative to tonal components in speech) [12]. In examining respiration, we used 1 feature of *Sound Pressure Level Decline* (SPLD – measure the ability to maintain intensity level). In examining articulation, we calculated 3 features including *DDK rate* (number of syllable vocalizations per second), *Robust Formant Periodicity Correlation* (RFPC – quantifies the accuracy of articulation), and *Spectral Distance Change Variation* (SDCV – quantifies the clarity of articulation) [10]. In examining prosody, we used 3 features including *fundamental frequency variation* (F0 SD), *intensity of voice variation* (Intensity SD), and *number of pauses* [10].

Two-sided Wilcoxon rank-sum test was performed to find differences between groups. To explore the extent of PD-related vocal impairment, we applied the *Wald task* decision-making theory to features' Gaussian kernel densities [13]. As a result, for all the features, the subjects were classified as PD (dysarthric speech performance), HC (intact speech performance), or "not sure" (indecisive situation – performance of wider norm of healthy people). The higher quantity of classifications as PD is associated with progression of PD vocal impairment, the higher quantity of classifications as HC is associated with healthy speech performance.

B. Identification of voice and speech disorders

In order to create reliable assessment in clinical practice, there is a need to test and find the optimal combination of acoustic measurements that gain a useful amount of information for separating early PD from HC. Therefore, we constructed a feature vector with 8 representative measurements including jitter, shimmer, NHR, HNR, SPLD, RFPC, SDCV, and F0 SD. To reduce dimensionality of the data and find the combination of acoustic measurements with the best classification accuracy, the exhaustive search of all possible combinations of features was performed using the method from statistical learning theory called s*upport vector machine* (SVM) [14]. The SVM classifier with Gaussian radial basis kernel was applied because it allows smooth, curved decision boundaries. On the basis of the decision boundary, the SVM classifier enables to build a predictive model which decides whether a subject belong to the PD or HC group. The choice of optimal SVM parameters was determined by an exhaustive search over a range of values. Cross-validation with the leave-one-out method was used to validate reproducibility (for possible new outcome samples) of SVM classifier; the 50 iteration was used for validation of each combination.

C. Relationships between acoustic features and severity of disease

In addition to speech data, for each of the PD patients, we have administered the *duration of disease* prior to recording, stage of disease according to the *Hoehn & Yahr* (HY) scale (disability scale comprised of stages 1 through 5, where 5 is most severe), and global motor impairment according to the *Unified Parkinson's Disease Rating Scale* (UPDRS) III (motor rating scale from 0 to 108, where 108 represents severe motor impairment). UPDRS III contains 27 items, each scored from 0 (no disability) to 4 (severe disability). We have also administered the three UPDRS III composite subscores including *bradykinesia* (sum of the UPDRS III items 23, 24, 25, 26), *postural instability and gait disorders* (PIGD - sum of UPDRS III items 27, 28, 29, 30), and *rigidity* (UPDRS III item 22). Subsequently, the Person product-moment correlation was used to find relationships between acoustic features and HY stage, duration of PD, UPDRS III score, and UPDRS III composite subscores.

IV. Results

Table 1 summarizes the comparison of subject parameters and speech parameters between Parkinsonian speakers and control group. The final data obtained were composed of 116 recordings (56 from the PD patients, and 60 from HC individuals). After applying the Wald task, we have found that 18/24 patients with PD indicate some form of vocal impairment that differs from the speech performance of the wider norm of healthy people. None of the HC speakers reached the specific dysarthric performance of patients with PD.

From all possible tested measurement combinations, 4 features including NHR, SPLD, RFPC, and F0 SD obtained the best classification score of 85.02%. The classification performance of the entire measurements subset was 81.67%. From individual measures, F0 SD obtained the best classification accuracy of 81.30%. The maximal correct overall classification accuracy was 76.40% using only sustained phonation, and 71.35% using only the DDK task.

In PD patients, there were no statistically significant correlations between the vocal parameters and the stage or duration of the disease. Accordingly, there were no statistically significant correlations between the vocal parameters and UPDRS III scores. However, the partial subscore of bradykinesia significantly correlated with the measure of articulation SDCV (R = -0.44, P < 0.05) and measures of phonation including jitter (R = 0.42, P < 0.05), NHR (R = 0.43, P < 0.05), and HNR (R = -0.44, P < 0.05). Admittedly, the subscore of rigidity correlated with HNR (R = -0.43, P < 0.05). There were no significant correlations between the vocal parameters and the UPDRS subscores of PIGD and speech.

Table 1: List of results of acoustic measures with mean±SD values and statistical comparisons between Parkinsonian and healthy groups.

	Subjects		Difference
	PD (n = 24)	HC (n = 22)	between PD and HC
Subject parameters			
Age (year)	60.92±11.24	58.73±14.61	P = .46
Male	n = 20	n = 15	n/a
Female	n = 4	n = 7	n/a
Duration of PD (month)	31.29±22.25	n/a	n/a
HY stage	2.19±0.48	n/a	n/a
UPDRS III score	17.42±7.14	n/a	n/a
Sustained phonation			
Phonation			
Jitter (%)	0.91±0.68	0.33±0.21	***P* < .001**
Shimmer (%)	8.57±4.60	3.25±1.57	***P* < .001**
NHR (-)	0.22±0.25	0.04±0.03	***P* < .001**
HNR (dB)	14.05±6.01	22.55±4.28	***P* < .001**
DDK task			
Respiration			
SPLD (1/s)	5.68±2.99	3.85±3.01	***P* < .05**
Articulation			
DDK rate (syll/s)	6.01±0.60	7.16±0.71	***P* < .001**
RFPC (-)	0.43±0.14	0.58±0.10	***P* < .001**
SDCV (-)	0.14±0.03	0.17±0.03	***P* < .01**
Running speech			
Prosody			
F0 SD (semitones)	1.52±0.43	2.62±0.75	***P* < .001**
Intensity SD (dB)	7.15.±1.42	8.66.±1.49	***P* < .001**
No. pauses (pause/s)	3.24.±0.85	3.83.±0.70	***P* < .01**

V. Discussion

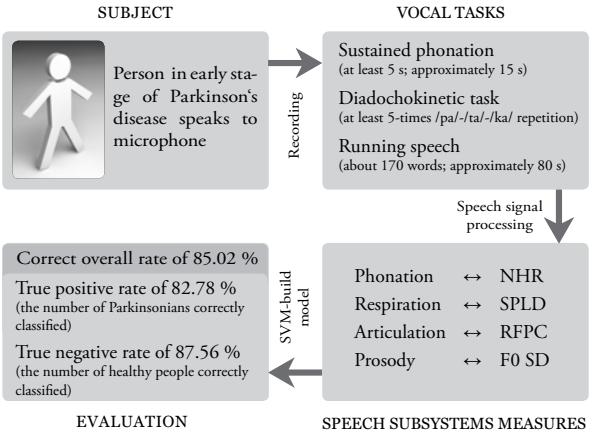
Fig. 1 summarizes the procedure and results of the two-minute vocal test that was employed to evaluate voice and speech disorders in a group of patients with unmedicated PD in comparison to HC people. For the sake of acoustic analysis, the measurements were designed as robust as possible with respect to a possible real-time automatic evaluation in a common acoustic environment and with the presence of contradictory factors such as individual differences in voice and speech. The acoustic measures were used as features for classification of probands into the PD and HC groups. Despite the limited number of speech samples, the best classification accuracy gains performance of 85% was reached using the combination of four measures, each of them representing deficits in one of the speech subsystems related to PD.

According to our results, the deficits in speech prosody appear to contain the greatest amount of information in assessment of early PD-related vocal impairment. Similarly, reduced melody in running speech captured by F0 SD measurement was found in other studies in PD patients, both treated and untreated with dopaminergic drugs [7, 15, 16]. On the other hand, several previous studies suggested that the most salient features of PD speech are related to phonatory and articulatory impairment [3, 4]. Indeed, our findings of increased values in jitter, shimmer, and NHR/HNR that may be clinically interpreted as hypophonia, voice hoarseness, and tremolo are in agreement with a previous report on untreated patients with PD [15]. However, in PD patients treated by dopaminergic drugs, only the jitter values were increased compared to controls while shimmer values were similar to those of controls, and the NHR/HNR findings were controversial [16].

VI. Conclusion

In conclusion, our newly designed configuration of vocal tests appears suitable for identification of voice and speech disorders in early stages of PD where it can accurately differentiate PD patients from HC. It consists of vocal tasks commonly used in most of the research studies examining PD-related voice and speech disorders [7]. Furthermore, the measurement methods can be

Figure 1: Schematic diagram depicting the recording of the PD patient's speech signals through the vocal test. Signals are calculated using speech signal processing algorithms and evaluated using the SVM-based model.



performed automatically with sufficient accuracy and without assistance of speech pathologist. In the future, when precise automatic assessment of boundaries between vowels and consonants etc. become feasible, the quick vocal test can be extended using new measurement methods such as for example novel acoustic measure of formant centralization ratio [17] which was proposed to robustly differentiate dysarthric from healthy speech.

ACKNOWLEDGMENT

This research was partly supported by the Czech Science Foundation, project GACR 102/08/H008, Czech Ministry of Health, projects NT 11331-6/2010 and NT 12288-5/2011, Grant Agency of the Czech Technical University in Prague, project SGS 10/180/OHK3/2T/13, and Czech Ministry of Education, projects MSM 0021620849 and MSM 6840770012.

REFERENCES

[1] H. Braak K. Del Tredici, U. Rüb, R. A. de Vos, E. N. Jansen Steur, and E. Braak, "Staging of brain pathology related to sporadic Parkinson's disease," *Neurobiol. Aging*., vol. 24, pp. 197-211, 2003.

[2] A. K. Ho, R. Iansek, C. Marigliani, J. Bradshaw, and S. Gates, "Speech impairment in large sample of patients with Parkinson's disease," *Behav. Neurol.*, vol. 11, pp. 131-137, 1998.

[3] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and coocurrence of vocal tract dysfunction in the speech of a large sample of Parkinson patients," *J. Speech. Hear. Disord.*, vol. 43, pp. 47-57, 1978.

[4] C. Ludlow, N. Connor, and C. Bassich, "Speech timing in Parkinson's and Huntington's Disease," *Brain. Lang.*, vol. 32, pp. 195-214, 1987.

[5] S. Skodda, and U. Schlegel, "Speech rate and rhythm in Parkinson's Disease," *Mov. Disord.*, vol. 23, pp. 958-992, 2008.

[6] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Hear. Res.*, vol. 12, pp. 246-269, 1969.

[7] A. M. Goberman, and C. Coelho, "Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy," *Neurorehab.*, vol. 17, pp. 237-246, 2002.

[8] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS)," *Eur. Arch. Otorhinolaryngol.*, vol. 258, pp. 77-82, 2001.

[9] S. Fletcher, "Time-by-count measurement of diadochokinetic syllable rate." *J. Speech Hear. Disord.*, vol. 15, pp. 757-762, 1972.

[10] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of voice and speech disorders in early untreated Parkinson's disease," *J. Acoust. Soc. Am.*, vol. 129, pp. 350-367, 2011.

[11] J. Rusz, R. Cmejla, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test," *Mov. Disord.*, 2011, in press.

[12] P. Boersma, and D. Weenink, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, pp. 341-345, 2001.

[13] A. Wald, "Sequential analysis," New York: Wiley, 1947.

[14] T. Hastie, R. Tibshirani, and J. H. Friedman, "The elements of statistical learning : data mining, inference, and prediction: with 200 full-colour illustrations," New York: Springer, 2001.

[15] F. J. Jimenez-Jimenez, J. Gamboa, A. Nieto, J. Guerrero, M. Orti-Pareja, J. A. Molina, E. Garcia-Albea, and I. Cobeta, "Acoustic voice analysis in untreated patients with Parkinson's disease," *Parkinsonism. Relat. D.*, vol. 3, pp. 111-116, 1997.

[16] J. Gamboa, F. J. Jimenez-Jimenez, A. Nieto, J. Montojo, M. Orti-Pareja, J. A. Molina, E. Garcia-Albea, and I. Cobeta, "Acoustic voice analysis in patients with Parkinson's disease treated with dopaminergic drugs," *J. Voice*, vol. 11, pp. 314-320, 1997.

[17] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," *J. Speech Lang. Hear. Res.*, vol. 53, pp. 114-125, 2010.

Session VI:
Devices

EXPERIENCES OF USING A DSP BASED DEVICE FOR VOCAL SIGNAL ANALYSIS

A. Palumbo[1], P. Veltri[1], B. Calabrese[1], P. Vizza[1], M. Cannataro[1],
A. Garozzo [2], N. Lombardo[2], F. Amato [1]

[1] Computer and Biomedical Engineering Laboratory, Department of Experimental and Clinical Medicine, University Magna Graecia, Catanzaro, Italy
[2] Otorhinolaryngoiatry Laboratory, Department of Experimental and Clinical Medicine, University Magna Graecia, Catanzaro, Italy

Abstract: **This paper presents the implementation of a DSP based acquisition and elaboration system for voice pathologies early identification. The proposed system performs real-time spectro-acoustic analysis of acquired voice samples and gives a visual feedback to alert about "potential" presence of anomalies in vocal robes. The prototype can also be used for rehabilitation purpose after medical treatment or surgery.**
Keywords: **Vocal tract pathologies, DSP technology, portable device, screening.**

I. INTRODUCTION

The identification of vocal tract diseases is carried out in clinical laboratories through invasive methods such as endoscopy. The analysis of vocal signal is a non-invasive method used for preliminary diagnosis and follow-up [1], and thus interests both medical doctors and engineers [2-7]. Vocal tract pathologies can be early identified by using signal voice analysis performed directly by patients or in a general doctor office.

We have been working on a project aimed to study and apply bioengineer techniques to vocal signals using otorhynolaryngoiatric expertise and experiences for the design and validation of a general purpose voice signal analysis system. Such a project aimed to: (i) improve vocal tract prevention providing home-care instruments, (ii) increase mass population screening, (iii) allow early vocal tract diseases detection and (iv) support rehabilitation phases improving follow up management 0[8-9].

This paper presents some experiences of using a portable digital signal processor (DSP)-based device for vocal signals acquisition and analysis, device resulting from the current status of the above described project. Tests have been performed on real data obtained from the University Magna Graecia otorhinolaryngoiatric laboratory, implementing time-frequency analysis of acquired voice samples. The tests show capabilities of the proposed device to give real-time feedbacks about relation between vocal signal anomalies and laryngeal pathologies.

II. METHODS

The designed device presents the following modules: (i) a microphone connected to the audio circuit and available on the board via 3.5 mm stereo jacks for vocal signal acquisition; (ii) A/D conversion; (iii) DSP processor for filtering and processing; (iv) output leds for patient feedback. It requires portability and usability with minimum weight and size; to this end the ADSP-BF537 Blackfin Processor (Analog Devices/Intel Micro Signal Architecture (MSA)) is used 0.

The DSP based device implements: (i) a preprocessing algorithm for data filtering, (ii) a voice signal feature extraction based on Short Time Fourier Transform (STFT), and (iii) the classification procedure based on the analysis of fundamental frequencies and (non) harmonic components.

Input information about patient sex and age have to be set before performing the analysis.

A. Preprocessing

The first stage of DSP processing consists of the digital filtering of vocal signals to improve the quality. The system is customized to implement different types of Butterworth filters (high-pass, low-pass, pass-band and stop-band) and specify the relative parameters (cut-off frequencies and orders) through push buttons available on the board. The advantage of Butterworth filters is their smooth and monotonically decreasing frequency response. More specifically, a Butterworth filter is an Infinite Impulse Response (IIR) filter. IIR filters, also known as recursive filters, operate on current and past input values and current and past output values.

Equation (1) defines the direct-form transfer function of the used IIR filter:

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

$$H(z) = \frac{b_0 + b_1 z^{-1} + \ldots + b_{Nb-1} z^{-N_b-1}}{1 + a_1 z^{-1} + \ldots + a_{N_{a-1}} z^{-N_a-1}} \quad (1)$$

where $b_0 \ldots b_{Nb-1}$ are the N_b forward coefficients, and $a_1 \ldots a_{Na-1}$ are the N_a reverse coefficient.

B. Voice signal feature extraction

The pathological voices are non-stationary signals because the frequency contents change over time. So, the Fourier Transform, that identifies the frequency components, does not allow to easily derive information about when and how these frequencies are actually present. For this reason, a time-frequency analysis is needed to detect temporal and spectral characteristics of the signal.

Short Time Fourier Transform (STFT) has been implemented on the DSP. To elaborate the STFT of the whole signal, it is divided into several blocks through a sliding window and then the Fast Fourier Transform (FFT) is applied to each data block to obtain the frequency contents.

The STFT is computed according to the following equation (2):

$$STFT[k,m] = \sum_{n=0}^{N-1} x_n w_{n,m} e^{-jk\frac{2\pi}{\Delta T}T_s n} \quad (2)$$

The discrete signal $x_n = x(nT_s)$, where $n=0, \ldots, N-1$ and T_s is the sampling period, is multiplied by the window function $w_{n,m}$, whose position varies in time, to obtain short-time segments. Specifically, the window function is defined as $w_{n,m} = w(nT_s + m\Delta T)$, where $m=0,1,\ldots, m\Delta T$ is the starting instant of the window and $\Delta T = NT_s$ is its duration. Then, the FFT is applied to each time segment.

C. Classification procedure

The spectrogram resulting from the STFT are represented in a numerical matrix S[N×M] where N is the number of frequencies and M is the number of samples; S[i,j] contains the power value of the jth sample corresponding to the ith frequency of the STFT of the vocal signal.

The procedure for the detection of pathological voice uses the information of fundamental frequency, first and second harmonics. The average power values of such frequencies are evaluated from S and compared with normality thresholds defined by the clinicians. Procedure evaluation also considers sub-harmonics and non-harmonics power values.

A procedure named *isPathological* is implemented on the DSP for pathological voice identification. The harmonic and sub-harmonic thresholds (Th0, Th1, Th2, SubTh0, SubTh1, SubTh2) need to be given as input to such procedure.

III. Results and Discussion

The designed device works in stand-alone mode or connected to a computer; in stand-alone mode, the device can be used by the patient directly that receives a visual feedback about the status of his voice. The PC-connected mode is mostly used by clinicians to configure the device to be furnished to patients and/or to gather information after it has been used.

The device has been tested on a data set gathered from the otorhinolaryngoiatric laboratories of University of Catanzaro, consisting of 31 patients of different ages and sex: 8 healthy (Subj1-Subj8) and 23 not healthy (Subj9-Subj31). Results are reported in Table 1. The system works correctly in all cases of healthy (normal signal); it returns a wrong diagnosis in 8 out of 23 not healthy subjects even if 4 out of 5 are related to Reinke's Edema pathology and 3 out of 6 to presence of nodules. This is mostly due to defect that this kind of pathologies require endoscopy analysis, i.e. even specialist visual and acoustic signal analysis is not sufficient for an early correct identification. Finally, results have been stored in PC-mode, and tests have been double checked by using standard spectrographic tools analysis, confirming the results of the portable device.

IV. Conclusion

In this paper a portable DSP-based system for the real-time acquisition and analysis of pathological voices is presented. The device is part of a project aiming to realize a mass population screening, early vocal tract diseases detection and voice rehabilitation. The device, that is associated with software tools for device configuration and signal analysis, is able to acquire, process the vocal signal and performs the analysis in time frequency domain. It can work in stand-alone mode, giving a visual LED-feedback to the patients about the voice, and in PC-based mode, showing the analysis results to clinicians for further studies. The prototype has been tested on a dataset of normophonic and pathological subjects and tested also on signals acquired from patients before and after medical treatments, showing its capability of being used for rehabilitation purposes.

Table. 1: Results of the Procedure

Medical Diagnosis	Subjects [Subj]	F0 [Hz]	Output Procedure
HEALTHY	Subj 1	288,303	Healthy
	Subj 2	143,659	Healthy
	Subj 3	221,747	Healthy
	Subj 4	176.334	Healthy
	Subj 5	214,156	Healthy
	Subj 6	217,218	Healthy
	Subj 7	264,244	Healthy
	Subj 8	218,609	Healthy
NOT HEALTHY (CHEP CHP)	Subj 9	123,32	Not Healthy
	Subj 10	227,937	Not Healthy
	Subj 11	143,657	Not Healthy
	Subj 12	182,148	Not Healthy
	Subj 13	89,724	Not Healthy
	Subj 14	125,391	Not Healthy
	Subj 15	225,571	Not Healthy
NOT HEALTHY (DISFUNCTIONAL DYSPHONIA)	Subj 16	210,337	Not Healthy
	Subj 17	286,723	Not Healthy
	Subj 18	249,66	**Healthy (wrong answer)**
NOT HEALTHY (REINKE'S EDEMA)	Subj 19	181,776	**Healthy (wrong answer)**
	Subj 20	129,32	Not Healthy
	Subj 21	133,917	**Healthy (wrong answer)**
	Subj 22	214,705	**Healthy (wrong answer)**
	Subj 23	116,803	**Healthy (wrong answer)**
NOT HEALTHY (NODULES)	Subj 24	277,491	Not Healthy
	Subj 25	228,539	**Healthy (wrong answer)**
	Subj 26	228,03	Not Healthy
	Subj 27	252,982	Not Healthy
	Subj 28	281,474	**Healthy (wrong answer)**
	Subj 29	218,369	**Healthy (wrong answer)**
NOT HEALTHY (POLYPS)	Subj 30	156,01	Not Healthy
	Subj 31	213,922	Not Healthy

REFERENCES

[1] J. Stemple, L. Glaze, B. Klapen, *"Clinical Voice Pathology"*, Fourth Edition, Plural Publishing, 2009.

[2] C. Manfredi, G. Peretti, *"A new insight into postsurgical objective voice quality evaluation: application to thyroplastic medialization"*, IEEE Trans Biomed Eng, 53(3) (2006) 442-451.

[3] E. S. Fonseca, R. C. Guido, P. R. Scalassara, C. D. Maciela, J. C. Pereira, *"Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders"*, Comput Biol Med, 37(4) (2007) 571-578.

[4] R. T. Ritchings, M. McGillion, C. J. Moore, *"Pathological voice quality assessment using artificial neural networks",* Med Eng Phys, 24(7-8) (2002) 561-564.

[5] B.S. Aghazadeh, H.K. Heris, *"Fuzzy logic based classification and assessment of pathological voice signals",* in Proc. of the Annual International IEEE EMBS Conference, (2009) 328-331.

[6] J. Wang, C. Jo, *"Vocal Folds Disorder Detection using Pattern Recognition Methods",* in Proc. of the 29th Annual International IEEE EMBS Conference, (2007) 3253-3256.

[7] C. Manfredi, T. Bruschi, A. Dallai, A. Ferri, P. Tortoli, *"Voice Quality Monitoring: a Portable Device Prototipe",* in Proc. of the 30th Annual International IEEE EMBS Conference, (2008) 997-1000.

[8] F. Amato, M. Cannataro, C. Cosentino, A. Garozzo, N. Lombardo, C. Manfredi, F. Montefusco, G. Tradigo, P. Veltri, *"Early Detection of Voice Disease via Web-Based System"*, Biomed Signal Process Control, 4(4) 2009 269-364.

[9] A. Palumbo, B. Calabrese, P. Vizza, N. Lombardo, A. Garozzo, M. Cannataro, F. Amato, P. Veltri, *"A Novel Portable Device for Laryngeal Pathologies Analysis and Classification"*, in: S. C. Mukhopadhyay, A. Lay-Ekuakille (Eds), Advances in Biomedical Sensing, Measurements, Instrumentation and Systems, Springer Verlag, 2009, pp.335-352.

[10] ADSP-BF537 Blackfin Processors Hardware Reference, Analog Devices, Revision 3.4, April 2009. Available at: http://www.analog.com.

ACCELERATION SENSOR MEASUREMENTS OF VIBRATIONS OF THE LARYNX IN PATIENTS WITH VOCAL FOLD ADDUCTION DEFICIENCIES

Wolfgang Wokurek*, Manfred Pützer†

*Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Deutschland
†Klinische Phonetik, Institut für Phonetik, Universität des Saarlandes, Saarbrücken, Deutschland
wokurek@ims.uni-stuttgart.de, puetzer@coli.uni-saarland.de

Abstract: **In this study we investigate voice productions of normal speakers and patients with varying vocal fold adduction deficiencies using a non-invasive method to measure spatial vibrations of the larynx. The current version ACCV4 of our acceleration sensor device was used. The study's primary goal is to find out whether lesions of vocal folds lead to additional spatial vibration modes compared to a non-disordered voice. Our results allow to assume that the composition of spatial vibra- tion modes at the skin over the cricothyroid ligament may depend on the symmetry of vocal fold movements.**

Keywords: **acceleration sensor, vocal fold adduction deficiencies**

I. INTRODUCTION

We use the current version ACCV4 of our acceleration sensor device to measure spatial vibrations of the skin of the neck covering the larynx. These vibrations are driven in part by the subglottal sound pressure and also indirectly by the vocal folds. We suspect that the movement of each vocal fold is passed on via their own arytenoid cartilage to the left or right part of the cricoid cartilage, respectively. The cricoid cartilage is situated at the lower part of the cricothyroid ligament [3]. Therefore, a path of vocal fold vibrations to the skin over this ligament can be assumend. Both the subglottal sound pressure and the symmetric vocal fold movements result in skin and sensor movements in the ventral and dorsal direction. These movements were attributed to the subglottal sound pressure alone in our previous studies [1], [4], [5], [6], [7]. Any deviations of vocal fold movements from symmetry could cause additional skin and sensor movements in lateral and/or cranial and caudal direction. We attempt to quantify the amount of asymmetry of the skin and sensor movement by an analysis of the spatial sensor movement.

II. METHODS

A. Acceleration Sensor

In this study the spatial vibrations of the current version ACCV4 of the acceleration sensor device are recorded simultaneously with the nasal and oral sound pressure signal captured separately through a Rothenberg mask.

Preceeding versions of the acceleration sensor de- vice ACCV4 were presented in [4], and [5]. The relevant aim in this study is its ability to track the spatial movement of the body tissue with a high bandwidth.

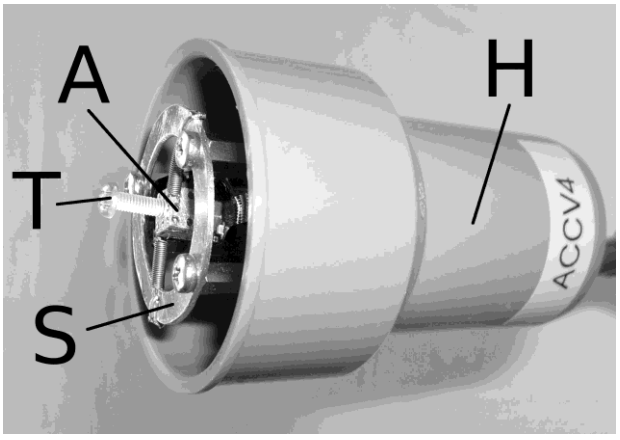


Figure 1: Acceleration sensor device

The acceleration sensor device is shown in Fig. 1. The tip T is mounted at the acceleration sensor A.

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Both are held by spiral springs in the suspension ring S. The suspension ring is glued to the handle H that also contains electrical connectors to conduct the analog signals to an external preamplifier by a cable.

The acceleration sensor A consists of three ADXL202E two axis microelectromechanical acceleration sensors that are glued to different planes of an aluminium cube. The tip T is a plastic screw fixed at the cube by a counter nut. The electronic components are soldered to a flexible printed circuit board (PCB). The force of the tip T to the neck is about

0.2 N – which proved to be strong enough to keep tissue contact but is hardly noticed by the spaker.

The arrangement of the ADXL202E chips tracks the acceleration along each spatial direction at two different points of the cube. Hence, this six signals are sufficient to compute the spatial vibrations of the body tissue.

B. Sensor Placement

The glottis is located in the larynx and separates the supraglottal from the subglottal cavity. It lies behind the thyroid cartilage. A soft tissue – the cricothyroid ligament – connects the lower end of the thyroid cartilage to the cricoid cartilage. The vocal fold vibration is passed on through the thyroid, arytenoid, cricoid cartilage, and the cricothyroid ligament to our sensor, respectively.

Figure 2: Sensor at neck

The cricothyroid ligament can be found by touching the larynx with a finger and searching for a small soft gap in the elsewhere hard larynx structure. The sensor tip T is placed perpendicular to the neck and pressed gently to the soft gap until the suspension ring S touches the skin as shown in Fig. 2. Now the speaker is asked to speak. The correct placement of the sensor is immediately seen in the amplitude display of the six accelerator channels. The amplitude of the two channels corresponding to the tip axis rise to high levels, the other four stay at low levels. In many cases this situation holds for several minutes. Sometimes the perpendicular position of the tip to the neck is lost and the signal amplitude distributes over more than two channels. In that case the session is paused and the sensor is arranged correctly again. In our recordings usually the sensor device was held by an assistant.

C. Speech sounds

Speech sounds are recorded via two electret micro- phones mounted in the oral and nasal section of a Rothenberg mask. In this study both sounds are added and used for labelling the short and long vowels. The lower part of the mask is visible in Fig. 2. The mask was held by the speaker.

D. Recordings

The recordings were made in a consultation room that was not sound treated. Eight channels were recorded simultaneously, six channels of the acceleration sensor as well as the oral and the nasal sound of the Rothenberg mask. The first order 5 kHz RC-lowpass recommended by the ADXL202E data sheet was implemented by analog hardware. All channels were digitized with a sampling rate of 48 kHz. The RME soundcard offers only AC coupling, hence no static ac- celeration signals like the gravitation vector are avail- able as a direction reference in the evaluation.

As speech material sustained vowels (e.g. [i:], [a:] [u:]) produced at the subject's normal pitch were used for this study.

E. Spatial analysis

A segment stable for about a second is located manu- ally in the sound of one of the sustained vowels. The temporal sampel indices n = 1, . . . , N correspond to that segment. To study the modes of vibrations of the sensor, the samples ai (n) of the six sensor channels are arranged in columns

$$a_i = (a_i(1), \ldots, a_i(N))^*, \quad i = 1, \ldots, 6 \qquad (1)$$

where $*$ denotes transposition. The columns are put together to form the N $\times$6 matrix of acceleration data

$$A = (a_1, \ldots, a_6) \quad . \qquad (2)$$

Each row of A may be viewed as a sample of the six dimensional vector valued sequence of acceleration measurements.

Whereas the mechanical sensor axes are not perfectly aligned parallel and perpendicular, no measurement of these deviation was done and hence no correction can be attempted. According to the data sheet of the sensor chip ADXL202E, the cross axis sensitivity of each sensor chip is ±2% or −34dB. It stems from axis misalignments and inherent sensor errors. Sensor assembly misalignments of 1 degree would result in additional cross sensitivity of about −39dB.

To find independent modes of vibration in the acceleration vector sequence A, the 6×6 correlation matrix is computed

$$R_A = A^* A \quad . \tag{3}$$

The eigen decomposition of this correlation matrix

$$R_A = V^* \Lambda V \tag{4}$$

results in the diagonal matrix

$$\Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_6) \tag{5}$$

of non-negative eigenvalues

$$\lambda_1 >= \lambda_2 >= \cdots >= \lambda_6 >= 0 \tag{6}$$

and in the orthogonal matrix

$$V = (v_1, v_2, \ldots, v_6) \tag{7}$$

containing the eigenvectors vi as columns.

According to Ineq.(6), the eigenvalues are arranged in the order of descending magnitude starting with the largest eigenvalue λ1 . The eigenvalue λ1 represents the energy of the major vibration mode. The direction of the major vibration mode is given by the corresponding eigenvector v1 . Similarly, the second vibration mode is given in energy and direction by λ2 and v2. Due to the symmetry of the correlation matrix, the eigenvectors corresponding to different eigen-values are always orthogonal. The case of equal or multiple eigenvalues with its associated subspace is not considered here further, since it never appeared in our measurements and it is very unlikely due to measurement noise.

The orientation of the major vibration mode is basically perpendicular to the skin at the neck, along the ventral and dorsal direction. The second mode vibrates along a line in the plane spanned by the lateral and the cranial and caudal direction. In order to quantify the symmetry of the sensor movement we propose the ratio between the energies of the major and the second vibration mode

$$\sigma = \lambda_1 / \lambda_2 \tag{8}$$

A large σ corresponds to a dominant major vibration

mode and a weak second vibration mode. In this situation the vibration of the cricothyroid ligament in lateral and/or cranial and caudal direction is weak

– a highly symmetric vibration. On the other hand a stronger vibration mode in lateral and/or the cranial and caudal direction reduces σ and corresponds to a more asymmetric vibration.

Since energy ratios may result in large figures the logarithmic decibel scale

$$\sigma_{dB} = 10 \log \sigma \tag{9}$$

is more familiar and often preferred. Both versions of the proposed symmetry measure σ and σdB will be shown in Tab.1.

F. Speakers

We investigated normal voices produced by two speakers with no known speaking or hearing problems as a control group. Additionally, three patients with varying vocal fold adduction deficiencies resulting from unilateral and bilateral paralysis of the recurrent nerve were considered [2]. This kind of pathology is a frequent cause of deficient vocal fold adduction. Patients compensate or do not use compensatory strategies for the adduction deficiency. Our three patients cover a wide range of physiological constellations. They were classified on the basis of the observed vocal fold adduction, judged from laryngoscopic and videostroboscopic recordings of their vocal folds during phonation by an experienced ENT physician. The clinical judgements were made during consultation.

III. RESULTS

The acceleration sensor device was previously used to get indirect access to the subglottal sound pressure and to measure the resonance parameters of the sub- glottal cavity [5]. It records the spatial components of the acceleration of its moving part. The analysis is based on the eigen decomposition of the correlation matrix of the acceleration vector. The projection to its main component was assumed to be driven mainly by the subglottal sound pressure. Now the strength of the second largest component is compared to the strength of the main component.

Table 1: Symmetry measure σ for normal adduction behaviour and different vocal fold adduction deficiencies.

	norm. voice	unilat. uncomp.	unilat. comp.	bilat. comp.
σ	15 23	4.7	12	133
σ_{dB}	12dB 14dB	7dB	11dB	21dB

Tab. 1 shows the resulting ratios of the underlying different vocal fold adduction behavior. Normal voices (norm. voice) having relative regular adduction behaviour (first column in Tab. 1) show a ratio of 15 and 23. The first vibration modes for these individuals are 12dB and 14dB stronger than their second modes.

Our first patient with uncompensated unilateral vocal fold paralysis (unilat. uncomp.; second column in Tab. 1) produces a much stronger second vibration mode which is only 7dB weaker than the first mode. This result may indicate non-symmetric vocal fold movements.

Our second patient with compensated unilateral vocal fold paralysis (unilat. uncomp.; third column in Tab. 1) offers results very closed to those of nor- mal voices. Consequently, compensation of vocal fold paralysis may restore symmetric vocal fold movements.

Finally, our third patient with compensated bilateral vocal fold paralysis (bilat. comp.; fourth column in Tab. 1) shows a very weak second vibration mode which may be caused by a high degree of symmetry in the vocal fold movement.

A closer look to the eigenvectors of our speakers confirms that the major mode vibrates, as conjectured, in the ventral and dorsal direction. The second vibration mode turns out to have always a component in the cranial and caudal direction. An additional lateral component in the second vibration mode is only seen with the second patient, not as expected with the first one.

IV. DISCUSSION

In the present study voice productions of normal speakers and patients with varying vocal fold adduction deficiencies were investigated. Instrumentally the current version ACCV4 of our acceleration sensor device was used. In extension to our previous approaches the spatial capabilities of the sensor were made use of. To measure the amount of symmetry of the vocal fold vibration the energy ratio of the first and second vibration mode was proposed and evaluated. It seems to mirror the symmetry condition of the vocal fold vibration despite of the underlying complex coupling path via arytenoid cartilage, cricoid cartilage, and the cricothyroid ligament.

V. CONCLUSIONS

Phonation behavior of patients with vocal fold adduction deficiencies resulting from unilateral and bi- lateral paralysis of the recurrent nerve show varying degrees of symmetry in their vocal fold vibration. The proposed symmetry measure of the energy ratio of the first and second vibration mode properly represents this situation. These observations encourage a further look at other phonation qualities to find out whether this symmetry measure is still applicable.

REFERENCES

[1] A. Madsack, S. Lulich, W. Wokurek, and G. Dogil. Subglottal resonances and vowel formant variability: A case study of high German monophthongs and Swabian diphthongs. LabPhon11, Wellington, Juli 2008.

[2] R. Sataloff. Professional Voice: The Science and Art of Clinical Care. Plural Publishing, San Diego, 2005.

[3] J. Wendler, W. Seidner, and U. Eysholdt. Lehrbuch der Phoniatrie und Pädaudiologie. Thieme, Stuttgart, 2005.

[4] W. Wokurek and A. Madsack. Messungen subglottaler Resonanzen mit Beschleunigungssen-soren. In 34. Jahrestagung für Akustik. DAGA, M¨arz 2008.

[5] W. Wokurek and A. Madsack. Comparison of manual and automated estimates of subglottal resonances. In Interspeech, Brighton, U.K., 2009.

[6] W. Wokurek and A. Madsack. Acceleration sensor based estimates of subglottal resonances: Short vs. long vowels. In Interspeech, Florence, Italy, 2011.

[7] W. Wokurek and M. Pützer. Acceleration sen- sor measurements of subglottal sound pressure for modal and breathy phonation quality. In MAVEBA09, Florence, December 2009.

VOICE MONITORING: TECHNICAL AND CLINICAL ASPECTS

Claudia Manfredi[1], Philippe H. Dejonckere[2]

[1]Department of Electronics and Telecommunications, Faculty of Engineering, Università degli Studi di Firenze, Via S.Marta 3, 50139 Firenze

[2]Catholic University of Leuven, Neurosciences, Exp. ORL, Belgium; Federal Institute of Occupational Diseases, Brussels, Belgium; Utrecht University UMC, AZU Heidelberglaan 100, F.02.504, NL 3584 CX Utrecht, The Netherlands

Abstract: **Occupational voice disorders are observed with increasing frequency in otolaryngological consultations. Devices have been developed that provide objective data on the way individuals use their voices throughout the day outside the clinic. However they do not provide information about the acoustic indexes of voice quality. A critical point is also the choice of the sensor.**

The device here proposed could be defined as a "portable laboratory" for voice analysis, its main advantage being the reliability of estimated parameters from both sustained vowels and running speech. A prototype has been set up on a DSP board and tested on short sustained vowels. Foreseen applications include: basic research, medico-legal, quantification of voice plasticity, vocal function exercises during rehabilitation, voice disorders, short term feedback in singing voice, etc.

Keywords : **Voice analysis, dosimeter, portable device, occupational voice disorders**

I. INTRODUCTION

The development of modern information telecommunication technology plays an increasingly important role in facilitating access to some diagnostic services, particularly in creating medical diagnostic applications small enough to fit into objects already in common use, such as cell phones.

Occupational voice disorders are observed with increasing frequency in otolaryngological consultations [1]. Speech therapists in voice clinical services rely on documenting information on therapy progress recording the examination/therapy session to diagnose the voice quality more precisely comparing the voice quality of the patient at the beginning, during and at the end of the therapy session and to review the evaluation later.

To this aim voice dosimeters and voice accumulators have been investigated, and suitable definitions of vocal load and dose have been given and applied to professional speakers and singers [2-9].

However few devices have been implemented, mostly based on a contact transducer (accelerometer) attached to the front part of the neck. A cable connects the accelerometer to the hardware module in a waist pack worn by patients. These devices provide data on the way how individuals use their voices throughout the day, outside the clinic, avoiding relying solely on subjective self-reports. In particular APM [10] records the total speaking time and sound level over a period of several hours. Quantitative measures of when, how long, how loud, and at what pitch the client vocalizes are obtained and a real-time feedback is provided, through a small vibrotactile unit. This information is very useful to identify those situations which might cause vocal fold damage. Other products implement similar voice quality parameters and indexes [11, 12]. Nevertheless they do not provide information about the acoustic indexes of voice quality. Another drawback with existing devices is the possible discomfort and embarrassment due to the contact transducer and the need of being returned to the clinic to download data into a PC for analysis using specific software. Moreover, a critical point is of course the correct wearing of the accelerometer that again could require clinical expertise.

The device here proposed differs from those above mentioned as is will be completely contact-less, the transducer being a small microphone. It could be defined as a "portable laboratory" for voice analysis, its main advantage being the reliability of estimated parameters of both sustained vowels and running speech and easy usage. Foreseen applications are:

- Research, to understand the early effects of fatigue on voice quality and/or the early mechanisms of vocal forcing.
- Medico-legal, by means of so-called "realistic provocation" test for patients that show a normal voice at the moment one examines them, but acknowledge a lot of voice symptoms during daily life.
- Quantification of voice plasticity in realistic conditions of use in voice professionals.
- Post-surgical monitoring and vocal function exercises during rehabilitation or for stuttering, dyslexia, psychogenic dysphonia, etc, where indirect interaction with the therapist could be more comfortable and effective.
- Occupational voice disorders (speakers, call center operators, teachers etc. with chronic voice over-use) to monitor how vocal folds react to the daily load and to

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

receive immediate feedback about possible risks. A long-term usage of the device could be foreseen over the days or weeks in subjects who are at risk for the development of voice pathologies, for easily available monitoring to be used by the voice clinician.

- Short term feedback in singing voice, while trying different vocal behaviours/voicing styles etc., e.g measuring the singer's formant and checking which one provides the best 'brilliance' to dominate the orchestra.

II. METHODS

A prototype has been set up on a DSP board that evaluates voice basic parameters and provides a LED/audio feedback that advices the patient for any abnormal vocal emission. The aim is to implement it on an object of common use, such as a cell phone, in order to overcome patient's distrust against medical devices. Data (audio files and parameters) could be saved on the device and possibly submitted to a PC for further analysis. This could be accomplished e.g. by means of MMS messages.

Voice quality indexes: On the prototype, voice quality analysis is based on the following indexes: fundamental frequency (F_0), along with its irregularities (Jitter, J, and Relative Average Perturbation, RAP), and hoarseness (Normalised Noise Energy, NNE). This is a subset of functions coming from a new user-friendly tool for voice analysis, named BioVoice [13] developed under Matlab R2009b, that can be easily extended to other relevant parameters related to vocal load.

Great attention is devoted to the selection of voiced/unvoiced frames as well as th F_0 estimation, as reliable estimates of other parameters depend upon it. The choice of the techniques adopted results from a detailed comparative analysis of F_0 extraction methods, with applications both to synthetic and real data in case of mild to severe dysphonia, showing enhanced performance against other approaches [14, 15].

The DSP prototype: The routines for F_0 and voice quality evaluation, developed under Matlab (release 2007b), were translated into C++ code (Microsoft Visual C++ 6.0) and optimised in order to run on the DSP board TMS320C6713, that is provided with a larger internal memory (192 kB) and faster clock (225 MHz) with respect to the one previously used [16]. The new DSP board allows for implementing many computations directly on a data buffer inside the internal memory. The buffer also allows for fewer transfers of data from external to internal memory. Moreover, floating point variables were implemented.

The board can work independently or connected to a laptop or PC (Fig. 1) for launching the debug and for showing on the monitor some plots as result of computations. The new DSP was also provided with the software required for audio signal recording through a microphone that must be connected to the MIC-IN input.

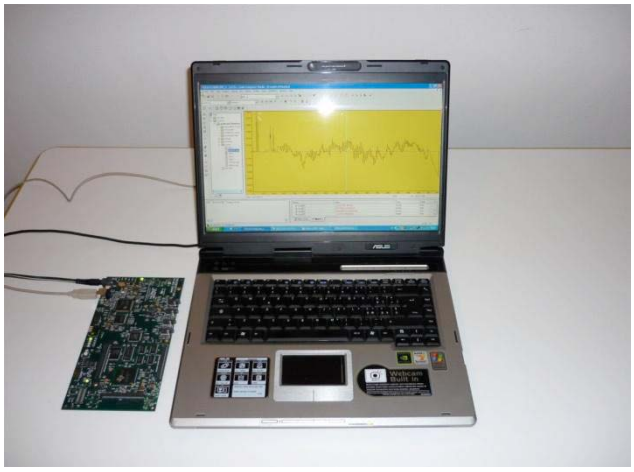


Figure 1: The DSP board connected to a laptop for further analysis and display of plots.

The microphone: A high quality voice recording should differentiate this system from existing ones, to allow for the evaluation of voice quality parameters and subtle differences induced by e.g. changing something in the voicing technique. Hence, the microphone plays a basic role in the device as it is required to be used in field. According to [17], if ambient noise or reverberation are a problem, as in this application, a head mounted omnidirectional microphone (that however reduces portability) or a directional microphone are suggested.

At present, the device works with a fixed distance of the mouth from a directional microphone. Recordings must be performed in a quiet environment. Users are carefully instructed and warned about these points, though a control has been implemented to test minimum amplitude, signal power and background noise requirements, in order to guarantee a satisfactory signal level while avoiding saturation. If such requirements are met, the signal analysis starts, according to the implemented algorithms. Otherwise, a devoted LED/audio alarm advices the patient that the recording must be repeated.

On the prototype the size of the data frame is limited to 2 s of recording, with F_s = 44 kHz sampling frequency, but of course longer frames are foreseen. More details can be found in [18].

As for any portable clinical device, before leaving the clinic to pursue her/his daily activities, the patient will receive instruction about how and when use the device. The clinician has to customize the device for each patient to elicit the audio/video alarm when a particular threshold, such as an irregularity value, is exceeded.

III. RESULTS

The new board has been tested on two sets of voice signals (sustained /a/ vowel). The first set consists of 40

pre-post surgical recordings with different degrees of hoarseness due to different pathologies (polyps, oedemas, cysts, tyroplastic prosthesis etc). The second set comes from healthy subjects recorded in non-protected environment, to test robustness against environmental noise.
The mean values of the parameters F_0, ANNE, J, and RAP were considered. The same computations were performed on both the DSP and the Matlab program running on a standard PC.
Table 1 shows some results from a subset of pathological signals (lines 1-8) and from male healthy subjects (lines 9-12). Only the results obtained from the DSP are reported, as they coincide up to the last digit with those obtained with Matlab. The computational time was 30-50 s on DSP, and 13 s on PC. Notice that the computational time could be greatly reduced if a reliable initial range of variation for F_0 is available that avoids a first step for F_0 estimation.

IV. Discussion

At present the proposed device is at a first stage of development, both as far as the implemented parameters and the hardware requirements are concerned. Adding other parameters of clinical relevance poses relatively simple problems, mainly concerning computational time, that could be solved with dedicated hardware or other techniques such as sending data to a server connected to a PC for visualization and further analysis with devoted tools (e.g. BioVoice). Parameters could include formants, spectrogram and PSD, as well as statistical results and plots on the whole recording period.
Sending data could be done through the GPRS/UMTS network (e.g. MMS messages) that should warrant for privacy, as the user could be identified through the telephone number available on a SIM card obtained only presenting a personal document and signing a legal document. The device could be provided with a HDD or memory card to store data and a USB connection to download data on a PC.
Moreover, an optimised version of the software could be developed to be downloaded as an application for mobile phones or i-phones.
The voice quality enhancement problem against environmental noise and/or simultaneous presence of other speakers rises more difficulties that could be partially solved applying e.g. blind source separation techniques or neural network algorithms to teach the phone to recognise the voice of the user against other voices or sounds. Another possibility could be that of applying spectral subtraction techniques, that would require having two microphones installed on the device. The first microphone should be very close to the mouth, as in usual mobile phones, while the second one should be mounted at a certain distance.
Also sound pressure level (SPL) should be taken into account. Some commercially available SPL meters do not need a calibration and will be investigated in future work
Another characteristics could be adding the possibility for the subject to indicate (by means of a button) relevant moments: the patient pushes on a button when he/she starts perceiving fatigue or burning throat, or other symptoms. Comparing voice quality before-during-after the button is pushed could give useful information in real time so that the patient could immediately react.
To keep the device user-friendly, at the output the user will be advised for abnormal phonation with intuitive audio/visual messages only. In-depth analysis is deferred to the complete analysis tool, available on a laptop or PC.

V. Conclusion

A DSP prototype is proposed for a contact-less portable device to be used by a patient in order to extract important parameters of vocal behaviour when pursuing normal activities. In addition to the important objective data the device provides, a real-time audio/video alarm is implemented, as a feedback tool to help patients to remind abusive vocal behaviours during routine daily activity and help the patient to learn how to modify vocal behaviour and achieve desired vocal function as defined by the clinician. This feature may enhance therapy carryover and expedite the patients rehabilitation process.
The proposed device could be useful for clinicians to monitor results of phonosurgery and to obtain objective acoustic data for statistical and scientific purposes avoiding expenses and time consumption to the patient under study.
Clinicians as well as speech therapists and psychiatrists could have benefits to obtain objective acoustic data for statistical and scientific purposes, avoiding a waste of money and time to the patient under study.
The possibility of making use of a simple and reliable self-monitoring tool, for non-expert users, with no restrictions on accessibility and logistics, will allow sensitising people on a still underestimated subject, such as the prevention of vocal apparatus pathologies

ACKNOWLEDGEMENTS

This work was supported by Ente Cassa di Risparmio di Firenze, under the project: “Interdisciplinary Laboratory for Biomedical Acoustics - LIAB”, n.2007.0754

REFERENCES

[1] P.H. Dejonckere (Editor), Occupational voice: care and cure, Kugler Pub., The Hague, The Netherlands, 2001.

Table 1 - F_0, Jitter, RAP, ANNE mean values and computational time with the DSP board. Slightly longer time (5"-10") is observed for female patients with respect to males, due to higher number of loops performed that is proportional to F_0 (usually higher in females).

File name	F_0 [Hz]	Jitter %	RAP %	ANNE [dB]	Comp. Time [s]
Pathol_male_1	104.4127	0.7395	0.3151	-17.0376	32.23
Pathol_ male_2	113.7185	0.4804	0.1961	-29.6496	33.49
Pathol_ male_3	156.7025	1.3709	0.4563	-27.6346	39.20
Pathol_ male_4	136.9295	0.6144	0.1039	-16.8766	36.31
Pathol_ female_1	202.0758	0.9030	0.0601	-28.6182	44.83
Pathol_ female_2	186.3836	0.3716	0.0833	-28.6481	43.42
Pathol_ female_3	184.5098	0.8682	0.0513	-29.2505	42.66
Pathol_ female_4	233.0304	0.5477	0.0584	-24.9314	50.07
FileTest1_a	116.5368	0.5320	0.0656	-23.7415	33.63
FileTest2_a	146.1851	0.2697	0.0000	-26.2802	37.49
FileTest3_a	114.4827	0.4110	0.0970	-24.7037	33.36
FileTest4_a	93.7563	0.7054	0.0942	-19.4262	31.45

[2] Zicker, J.E., Tompkins, W.J. Rubow, R.T. Abbs, J.H. (1980), A portable microprocessor-based biofeedback Training device, IEEE Trans. Biomed. Eng., 27, 509-515.

[3] Szabo, A., Hammarberg, B., Håkansson, A., Södersten, M., (2001), A voice accumulator device: evaluation based on studio and field recordings. Logoped. Phoniatr. Vocol. 26, 102-117.

[4] Cheyne, H.A., Hanson, H.M. , Genereux, R.P., Stevens, K.N. , Hillman, R.E., (2003), Development and testing of a portable vocal accumulator, J. Speech Lang. Hear. Res. 46(6),1457-67.

[5] R.E. Hillman , J.T. Heaton , A. Masaki , S.M. Zeitels , H.A. Cheyne . (2006), Ambulatory monitoring of disordered voices, Annals of Otology, Rhinology, and Laryngology 115, 795-801.

[6] Titze, I. R., Švec, J. G., Popolo, P. S., (2003), Vocal Dose Measures: Quantifying Accumulated Vibration Exposure in Vocal Fold Tissues, J Speech Lang Hear Res, 46, 919 – 932

[7] Carroll, T., Nix, J., Hunter, E., Emerich, K., Titze, I., Abaza, M., (2006), Objective measurement of vocal fatigue in classical singers: A vocal dosimetry pilot study, Otolaryngol. - Head Neck Surg. **135**, 595-602.

[8] Kelchner, L. N., Toner, M. M., Lee, L., (2006), Effects of Prolonged Loud Reading on Normal Adolescent Male Voices, Lang Speech Hear Serv Sch, 37, 96 - 103.

[9] J.Nix, J. Svec, A. Laukkanen, I. Titze. Protocol Challenges for On-the-Job Voice Dosimetry of teachers in the United States and Finland. Journal of Voice 21 (2007): 385-396.

[10] www.kayelemetrics.com/Product20Info/3200/3200.htm

[11] www.wevosys.com

[12] www.sonvox.com

[13] Manfredi, C., Bocchi, L., Cantarella, G., (2009), A multipurpose user-friendly tool for voice analysis: application to pathological adult voices, Biomed. Signal Proc. and Control, **4**, 212-220 .

[14] Manfredi, C., D'Aniello, M., Bruscaglioni, P., Ismaelli, A., (2000), A comparative analysis of fundamental frequency estimation methods with application to pathological voices, Med. Eng. Phys., **22**, 135-147.

[15] Manfredi, C., Giordano, A., Schoentgen, J., Fraj, S., Bocchi, L., Dejonckere, P.H., (2011), Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools, Biom. Signal Proc. and Control (in print).

[16] Manfredi, C., Bruschi, T., Dallai, A., Ferri, A., Tortoli, P., Calisti, M., (2008), Voice Quality Monitoring: a Portable Device Prototype", Proc. 30th EMBS Conf., Vancouver, CA, 997-1000.

[17] Granqvist, S., Švec, J., (2009), Microphones and Room Acoustics and Their Influence on Voice Signals, 8th Europ. Voice Conf., PEVOC8, Dresden, Germany.

[18] C.Manfredi, G.Cantarella, "A contact-less portable device for voice quality monitoring", Riv. Ital. Acust., vol. 35, n.1, p.53-57, 2011.

AUTHOR INDEX

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 7th international workshop : August 25-27, 2011*, ISBN 978-88-6655-009-9 (print) ISBN 978-88-6655-011-2 (online)

Finito di stampare presso
Grafiche Cappelli Srl - Osmannoro (FI)